PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 :
C12N 15/31, C07K 15/00
G01N 33/569, A61K 39/00

A1

(11) International Publication Number: WO 94/09141

(43) International Publication Date: 28 April 1994 (28.04.94)

(21) International Application Number: PCT/US93/09635

(22) International Filing Date: 8 October 1993 (08.10.93)

(30) Priority data:
07/958,683 9 October 1992 (09.10.92) US

(71) Applicant: THE GOVERNMENT OF THE UNITED STATES OF AMERICA as represented by THE DEPARTMENT OF HEALTH AND HUMAN SERVICES [US/US]; National Institutes of Health, Box OTT, Bethesda, MD 20892 (US).

(72) Inventors: KOVACS, Joseph, A. ; 11936 Goya Drive, Potomac, MD 20854 (US). ANGUS, C., William ; 6133 Fieldcrest Court, Frederick, MD 21701 (US). POWEL, Francoise ; 11420 Falcon Bridge Court, Beltsville, MD 20705 (US). EDMAN, Jeffrey, C. ; 126 Marion Avenue, Mill Valley, CA 94941-2617 (US).

(74) Agents: MURPHY, Gerald, M., Jr. et al.; Birch, Stewart Kolasch & Birch, P.O. Box 747, Falls Church, VA 22046-0747 (US).

(81) Designated States: AU, CA, JP, European patent (AT, BE CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: GENES THAT ENCODE A SURFACE PROTEIN OF P. CARINI'I

(57) Abstract

Seven unique genes encoding the major surface glycoprotein of rat *P. carinii*, which is related to the major surface antigen of *P. carinii* which is a life-threatening opportunistic pathogen in HIV-infected patients, have been cloned and sequenced. Genes encoding for the major surface glycoprotein of human *P. carinii* and vaccines containing the human *P. carinni* antigen can be prepared to prevent or control *P. carinii* infection.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria
AU	Australia
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CS	Czechoslovakia
CZ	Czech Republic
DE	Germany
DK	Denmark
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakhstan
LI	Liechtenstein
LK	Sri Lanka
LU	Luxembourg
LV	Latvia
MC	Monaco
MG	Madagascar
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SI	Slovenia
SK	Slovak Republic
SN	Senegal
TD	Chad
TG	Togo
UA	Ukraine
US	United States of America
UZ	Uzbekistan
VN	Viet Nam

GENES THAT ENCODE A SURFACE PROTEIN OF P. CARINII

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to the cloning of the major surface antigen of Pneumocystis carinii, a life threatening opportunistic pathogen in HIV-infected patients and the use of that antigen as a vaccine to prevent or control P. carinii infection.

Description of Related Art

The AIDS epidemic was heralded by the occurrence of Pneumocystis carinii pneumonia, a life-threatening opportunistic disease, in patients with no previously identified immunodeficiency (1,2). Subsequently, the number of cases of P. carinii pneumonia increased dramatically as human immunodeficiency virus (HIV) infection became wide-spread and the virus progressively impaired the immune system of infected patients. Over the past 5 years, important advances in the diagnosis, treatment, and prevention of P. carinii pneumonia have resulted in a decline in the frequency of P. carinii pneumonia, as well as an improvement in survival (3).

Despite these important clinical advances, the immunopathogenesis of P. carinii pneumonia is poorly understood. Although long considered a protozoan, recent molecular biologic studies have shown P. carinii to be a

member of the fungi (4-7). The major surface antigen of P. carinii is a mannose-rich glycoprotein of approximately 110,000 to 120,000 MW under reducing and denaturing conditions, with a native M_r (molecular weight) of over 300,000 (8-12). P. carinii isolated from different mammalian species contain similar but antigenically distinct proteins (8,9,13). The present inventors have recently purified and characterized the major surface protein of both rat (gp116) and human (gp95) P. carinii, and have demonstrated that about 10% of the M_r is accounted for by N-linked carbohydrates, with distinct carbohydrate profiles for the two proteins (8). Recent studies have suggested that gp116 is important in organism-host cell binding, possibly through interactions with fibronectin (14), mannose-binding protein (15), or surfactant protein A (16). Passive immunization studies with a monoclonal antibody directed against a conserved epitope of this antigen have demonstrated partial protection against P. carinii pneumonia in rats and ferrets (17). The major surface antigen thus appears to play a role not only in host-organism interactions, but also in host defense mechanisms.

SUMMARY OF THE INVENTION

Based on Southern blot studies using chromosomal or restricted DNA, the major surface glycoproteins of P. carinii have been found to be the products of a multicopy family of genes. The predicted protein has a MW of approximately 123,000, is relatively rich in cysteine residues (5.5%) that are very strongly conserved, and contains a well-conserved hydrophobic region at the carboxy terminus. The presence of multiple related genes encoding the major surface glycoprotein of P. carinii suggests that antigenic variation is an important mechanism for evading host defenses.

The present inventors have isolated and sequenced the DNA (and deduced the corresponding amino acid sequences)

for seven unique genes, each of which encodes a major surface glycoprotein of rat P. carinii. These genes are related to the corresponding genes in human P. carinii. Seven cDNA clones, PC3, PC5, PC14, GP3, GP22, GP46 and GP14, encoding gp116, were isolated and the sequences obtained suggest that gp116 is, in fact, a heterogeneous mixture of proteins encoded by multiple related genes. This should enable the preparation of the corresponding DNA in P. carinii which infects humans and the corresponding polypeptides, thus permitting the development of a vaccine based on the major surface antigen of P. carinii strains which infect P. carinii.

Accordingly, it is an object of the present invention to provide and isolate a DNA molecule encoding a mammalian Pneumocystis carinii major surface glycoprotein or allelic variations thereof. It is also an object of the invention to provide a DNA molecule encoding the gene for the major surface glycoprotein of P. carinii as shown in Figure 1b. It is a further object of the invention to provide a DNA molecule encoding all or a portion of the gene for the major surface glycoprotein of P. carinii in a cDNA clone including the clones of PC3, PC5, PC14, GP3, GP22, GP46 and GP14.

It is an additional object of the invention to provide DNA molecules which encode human P. carinii major surface glycoprotein or allelic variations thereof.

It is another object of the invention to provide a method of obtaining a DNA molecule encoding a mammalian P. carinii major surface glycoprotein which comprises screening a cDNA expression library of P. carinii with an antibody to said major surface glycoprotein to identify positive clones encoding gp116 and using at least one of said clones or an oligonucleotide probe based on said clones to reveal the presence of multiple genes encoding for said major surface glycoprotein.

It is a further object of the invention to provide a mammalian Pneumocystis carinii major surface glycoprotein having the amino acid sequence as shown in Figure 1b.

It is an additional object of the invention to provide a mammalian Pneumocystis carinii major surface glycoprotein produced from the expression of a DNA sequence which is a composite (or consensus sequence) of multiple genes which encode said major surface glycoproteins.

It is a further object of the invention to provide a human Pneumocystis carinii major surface glycoprotein produced from the expression of a DNA sequence which is a composite (or consensus sequence) of multiple genes which encode said major surface glycoprotein.

It is another object of the invention to provide a vaccine comprising a therapeutically effective amount of a mammaian Pneumocystis carinii major surface glycoprotein or a polypeptide derived therefrom capable of eliciting an immune response to said glycoprotein, and pharmaceutically acceptable parenteral vehicle.

It is also an object of the invention to provide a DNA molecule encoding a mammalian Pneumocystis carinii major surface glycoprotein which is a composite (or consensus sequences) of multiple genes which encode said major surface glycoprotein.

Further scope of the applicability of the present invention will become apparent from the detailed description and drawings provided below. However, it should be understood that the detailed description and specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only since various changes and modifications within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and other objects, features, and advantages of the present invention will be better understood from the

following detailed descriptions taken in conjunction with the accompanying drawings, all of which are given by way of illustration only, and are not limitative of the present invention, in which:

Figure 1A-1B. Alignment of the deduced amino acid sequences SEQ ID NO: 1 through SEQ ID NO: 7 represent 7 homologous clones encoding the major surface glycoprotein of rat P. carinii. Alignment was performed by the Clustal program of PC-Gene (IntelliGenetics, Inc.). Cysteine residues are identified in bold. Potential glycosylation sites are underlined. The peptides sequenced directly are shown above the alignment. An * indicates that a residue is conserved among all clones that overlap in that region.

Figure 2. Immunoblots demonstrating reactivity of anti-peptide antibodies with the major surface glycoprotein of rat P. carinii. Lanes 1, 3, 6, and 8, whole-organism extract; lanes 2, 4, 5, and 7, lyticase-solubilized proteins. Lanes 1, 2, 5, and 6, pre-immune sera (1:100); lanes 3 and 4, hyperimmune serum (1:100) following immunization with $GP3_{446-460}$; lanes 7 and 8, hyperimmune serum (1:100) following immunization with $PC5_{365-379}$. Reactivity specifically with the major surface glycoprotein (M_r=116,000) is seen with both hyperimmune sera. Lyticase treatment solubilizes the major surface glycoprotein, but results in a loss in apparent M_r of about 10%. Samples were run on a gradient gel (8% to 16%) prior to transfer to nitrocellulose. Migration of molecular weight markers is indicated on the left.

Figure 3A. Southern blot of P. carinii DNA (20 μg/lane) digested with Nde 1 (first lane of each pair) and Eco R1 (second lane of each pair) and subsequently probed with MSG1 (common sequence), MSG2 (GP3-specific), MSG3 (GP14-specific), or DHPS1. Standards in kilobases are indicated on the left. P. carinii DNA was obtained from a single infected rat. None of the regions from which the oligonucleotides were derived contain Eco R1 or Nde 1 sites. All oligonucleotides were labeled at the same time,

and approximately equal numbers of counts were added for each probe. The first two lanes were exposed for 4 hours, and the remaining lanes for 48 hours. The presence of multiple bands in the first two lanes demonstrates that multiple copies of these genes are present. Fewer bands are seen with the oligonucleotides specific for GP3 or GP14, but all bands correspond to those seen with the common oligonucleotide. No hybridization with MSG1 was seen with rat DNA (blot not shown). Hybridization with DHPS1, derived from P. carinii dihydropteroate synthase, demonstrates the intensity of reactivity with a presumed single-copy gene.

Figure 3B. Southern blot of P. carinii chromosomes from 5 isolates (lanes 1 to 5) and Saccharomyces cerevisiae chromosomes (lane 6) separated by transverse alternating field electrophoresis (28) and probed with PC5, demonstrating hybridization with multiple P. carinii chromosomes in all isolates. MW, based on S.cerevisiae chromosomes, is indicated on the right.

Figure 3C. Northern blot of total RNA extracted from 3T3 cells (10 μg, lane 1) or 5 P. carinii isolates (5-10 μg, lanes 2 to 6) probed with PC5. Hybridization to an approximately 4000 bp transcript is seen in P. carinii lanes. Migration of rRNA is indicated on the right.

DETAILED DESCRIPTION OF THE INVENTION

The following detailed description of the invention is provided to aid those skilled in the art in practicing the present invention. Even so, the following detailed description of the invention should not be construed to unduly limit the present invention, as modifications and variations in the embodiments herein discussed may be made by those of ordinary skill in the art without departing from the spirit or scope of the present inventive discovery.

The contents of each of the literature citations in the present application are herein incorporated by reference in their entirety.

The nucleotide sequence of clone PC3 (SEQ ID NO: 1) encodes for a portion of the coding sequence for the major surface glycoprotein of rat P. carinii.

The nucleotide sequence of clone PC5 (SEQ ID NO: 2) encodes for a portion of the coding sequence for the major surface glycoprotein of rat P. carinii.

The nucleotide sequence of clone PC14 (SEQ ID NO: 3) encodes for a portion of the coding sequence for the major surface glycoprotein of rat P. carinii.

The nucleotide sequence of clone GP3 (SEQ ID NO: 4) encodes for a protein similar to the original gp116 clones and having a molecular weight of 104,048.

The nucleotide sequence of clone GP46 (SEQ ID NO: 5) encodes for a portion of the major surface glycoprotein of rat P. carinii.

The nucleotide sequence of clone GP22 (SEQ ID NO: 6) encodes for a portion of the major surface glycoprotein of rat P. carinii.

The nucleotide sequence of clone GP14 (SEQ ID NO: 7) encodes for a portion of the major surface glycoprotein of rat P. carinii.

DNA (SEQ ID NO: 8) and inferred amino acid sequence (SEQ ID NO: 9) illustrate one gene of the major surface glycoprotein of P. carinii. The DNA sequence, which was determined from both strands, is a composite of the original GP3 clone (SEQ ID NO: 4) (nucleotides 626 to 3521) and the 5′ fragment (1 to 722) that was determined by PCR. Primers used in PCR to identify the 5′ end of the sequence are underlined once, and the 5′ end of the original clone is underlined twice. The 5′ fragment was missing the first nine nucleotides of the 5′ primer. The polyadenylation signal is shown in bold.

MATERIALS AND METHODS

Materials

Restriction enzymes were obtained from New England Biolabs (Beverly, MA). Other enzymes or kits were obtained from Stratagene (La Jolla, CA), Boehringer Mannheim (Indianapolis, IN), or InVitrogen (San Diego, CA). Polymerase chain reaction (PCR[1]) studies were performed with a DNA thermocycler (Perkin-Elmer/Cetus) using reagents obtained from Perkin-Elmer/Cetus. Radiolabeled chemicals were obtained from New England Nuclear-DuPont (Boston, MA). Sequenase 2 was obtained from United States Biochemical (Cleveland, OH). Oligonucleotides were synthesized on a Cyclone-Plus DNA synthesizer (Milligen Biosearch, Burlington, MA) using reagents obtained from Milligen. Hybond-N+ was obtained from Amersham (Chicago).

P. carinii organisms. Organisms were obtained from immunosuppressed rats and partially purified by Ficoll-Hypaque density gradient centrifugation as described (18).

P. carinii libraries. Construction of a P. carinii cDNA library in λ ZAP has been described (4). A second library was constructed in a similar fashion using oligo-dT-selected mRNA and subcloning into a modified λ ZAP vector (19), YcDE11, which contained sequences necessary for Saccharomyces cerevisiae replication and expression (Edman, J.C., unpublished observations). Both libraries were constructed from RNA pooled from three P. carinii preparations.

General Methods

Screening of libraries. Antibody screening was performed by described techniques (20) on approximately 50,000 phage following induction with isopropyl-β-D-thiogalactopyranoside (10 mM) using serum (1:1000) from a rat immunized with rat P. carinii (18). Positive clones were plaque purified, and clones encoding the major surface glycoprotein were identified by the antibody elution technique (21). Briefly, approximately 5,000 phage were

plated with BB4 cells on NZCYM agar; after 3-4 hours growth at 42°C, plates were overlaid with nitrocellulose that had been soaked in 10mM isopropyl-β-D-thiogalactopyranoside, and incubated overnight. Filters were blocked, incubated overnight with hyperimmune rat serum (1:1000), and washed. Reactive antibodies were eluted using 5 mM glycine-HCI, pH 2.3, 150 mM NaCl, 0.5% Triton-X 100, and 100 μg/ml bovine serum albumin. After neutralization, eluted antibodies reactive with the major surface glycoprotein were identified by the immunoblot technique using P. carinii antigens as described (18). For screening with DNA, probes were labeled with [α-^{32}P]-dCTP using the random priming method (22). Hybridization was performed overnight at 65°C in 6x SSPE/1%SDS/10x Denhardt's solution (1x SSPE is 0.15 M NaCl, 10mM NaH_2PO_4, 1 mM EDTA-Na_2,pH 7.4; 1x Denhardt's is 0.02% polyvinylpyrrolidone, 0.02% Ficoll, 0.02% bovine serum albumin). Filters were washed at 65°C in 0.5xSSPE/0.1% SDS. Positive clones were plaque purified, and plasmids (pBluescript plus insert) were rescued according to the manufacturer's instructions (Stratagene) (19). Inserts were sequenced directly from plasmid using the Sanger dideoxy chain termination method (23) either in the inventors' laboratory using the Sequenase 2.0 kit or commercially (Lofstrand, Gaithersburg, MD).

Polymerase chain reaction (PCR): The 5' region of GP3 was determined by PCR. To identify the 5' end of the mRNA, oligonucleotide JK58, complementary to positions 306 to 325 of PC3 (SEQ ID NO: 10) (TTAACCGGCCGTGCCATTGC), which includes the putative initiation codon, was used as a template for reverse transcription (24), after which the cDNA was tailed with terminal transferase and dGTP, amplified by PCR as described (25), using primer JK58 and a 1:10 ratio of modified primers ANC SEQ ID NO: 11 (GACTGCATGCGGAAGCTTGGATCCCCCCCCCCCCCC) and AN (SEQ ID NO: 12) (GACTGCATGCGGAAGCTTGGATCC), subcloned into pCR1000 (Invitrogen), and sequenced. The region 5' to GP3 was then determined by reverse transcription followed by PCR (24),

using a 5' primer corresponding to the previously determined first 20 bases of the mRNA (SEQ ID NO: 13) (TTTTTCTAATAGACGATATG), and a 3' primer complementary to positions 77 to 96 of GP3 (SEQ ID NO: 14) (GATCTCCACATGTTTTAGCA), subcloning as above and sequencing.

Southern and Northern blots. For Southern blots, P. carinii DNA (20 μg/lane) digested with Eco R1, or Nde 1 was probed with the following oligonucleotides that had been labeled with [γ-^{32}P]ATP using T4 polynucleotide kinase (26):

MSG1:GCAGAACTTGAGTCGGAATGTTT[C,T]TATTTA (SEQ ID NO: 15);

MSG2:AAAATATCTTCCACGATGTCTTTATCCTAA (SEQ ID NO: 16);

MSG3:GAAAATAAAGATAAGAGATACCTTCCAAAG (SEQ ID NO: 17); and

DHPS1:

TTGATCACGATATTAAGCCAGTTTTGCCAT (SEQ ID NO: 18). MSG1, which corresponds to nucleotides 1346 to 1375 of GP3, is well conserved among the overlapping clones. MSG2 (1573 to 1602 of GP3) and MSG3 (223 to 252 of GP14) are based on regions of PG3 and GP14 that are poorly conserved in other clones. DHPS1 is complementary to 1897 to 1926 of the P. carinii *fas* gene, which encodes P. carinii dihydropteroate synthase (27). None of the oligonucleotides contained Eco R1 or Nde 1 sites. For pulse-field gels, P. carinii chromosomes were separated by transverse alternating field electrophoresis as described (28) and probed with [^{32}P]-labeled PC5 or MSG1. For northern blots, RNA was extracted using an RNA isolation kit (Stratagene) according to the manufacturer's instructions; 5-10 μg total RNA was separated by formaldehyde/agarose gel electrophoresis and probed with [^{32}P]-labeled PC5. All blots were transferred to Hybond-N+. Blots probed with PC5 were prehybridized overnight in 6x SSPE/1%SDS/10x Denhardt's solution, hybridized overnight at 65°C with [^{32}P]-labeled PC5 (55°C for chromosome blots), then washed twice for 5 min. at room temperature in 2xSSPE/0.1% SDS followed by two washes for 20 min. at the hybridization temperature in 0.5xSSPE/0.1%SDS or, for chromosomal blots, 0.1xSSPE/0.1%/SDS. Blots probed with oligonucleotides were

prehybridized overnight in 2x SSPE/0.5%SDS/5x Denhardt's solution/0.5 μg/ml sonicated and denatured salmon sperm DNA, hybridized overnight at 60°C with [^{32}P]-labeled oligonucleotide, and washed three times at 60°C for 30 minutes each in 2xSSPE/0.5% SDS.

Peptide sequencing. Peptide sequencing was performed (Harvard μChem, Cambridge, MA) on peptides of gp116 following treatment of purified gp116 (8) with endoproteinase LysC and separation by narrow-bore reverse phase HPLC using previously described techniques (29). Peptide 1 was selected for sequencing by analyzing the predicted peptide sequences originating between the first two predicted methionines as follows: first, peptide retention prediction suggested such peptides would be retained predominantly in the first quarter of the chromatogram. Second, greater than 70% of the sequences lacked tryptophan or tyrosine and, thus, would be deficient in UV absorbance at 277 nM. On the basis of these two criteria, appropriate fractions were screened by electrospray ionization mass spectrometric analysis for a molecular mass matching a sequence from the desired region (29).

Anti-peptide antibodies. The following peptides were synthesized by Peninsula Laboratories (Belmont, CA): $GP3_{446-460}$ (SEQ. I.D. NO.:9, amino acids 446-460) (Glu-Leu-Lys-Gly-Lys-Leu-Gly-His-Val-Arg-Phe-Tyr-Ser-Asp-Pro), which corresponds to amino acid residues 446-460 of GP3 and 453 to 467 of PC3; and $PC5_{365-379}$ (SEQ. I.D. NO.:19) (Glu-Leu-Arg-Gly-Asn-Leu-Gly-Leu-Val-Arg-Phe-Tyr-Ser-Asp-Pro), which corresponds to 365 to 379 of PC5. Peptides (10 mg) were commercially coupled (Peninsula Laboratories) to KLH (50 mg) (30), and two rabbits were immunized (Lofstrand) with 0.5 to 1.25 mg of each peptide conjugate every two weeks for 10 weeks, using complete (first dose) or incomplete (remaining doses) Freund's adjuvant. Immunoblots against whole organism extracts or lyticase-solubilized gp116 were performed as previously described (8).

Computer analysis. DNA and protein analyses were performed using either the PC-Gene (Intelligenetics) or MacVector (IBI) analysis programs.

Molecular weight (M_r) Determinations. The M_r for purified major surface glycoprotein was determined for the native protein by a sizing column used on an HPLC, and for the reduced and denaturated protein by SDS-polyacrylamide gel electrophoresis. For the proteins encoded by the cloned genes, the M_r was determined by the computer program MacVector, based on the amino acid sequences of the predicted proteins.

Gene composite. The composite was generated by alignment of the sequence of the original GP3 clone and the sequence of the PCR-generated fragment corresponding to the 5' end of the gene. Nucleotides 626 to 722 of the PCR fragment were found to be identical to nucleotides 1 to 96 of the original GP3 clone. This allowed appending nucleotides 1 to 625 of the PCR fragment to the 5' end of GP3, to generate the composite full-length clone.

Consensus sequence synthesized gene. A consensus sequence can be generated by computer alignment of the proteins encoded by each gene from the multiple clones. A clone containing a synthetic construction and representing the consensus sequences, or regions that contain some of the consensus sequences, could then be derived by molecular biologic techniques, for example by replacing regions in one of the clones with consensus sequences, or by using site-directed mutagenesis (39).

RESULTS

Identification of genes encoding the major surface glycoprotein. Multiple clones were identified by immunoscreening a rat *P. carinii* cDNA library using rat serum generated against whole rat *P. carinii* (18). Clones reactive with polyclonal serum were evaluated by the antibody elution technique (21) to identify those potentially encoding for gp116. These clones cross-

hybridized by Southern hybridization; however, it was not possible to align the clones by restriction mapping. Three of these clones (PC3, PC5, and PC14) were sequenced in their entirety and contained open reading frames encoding three closely related but distinct proteins. Although none of the clones contained the complete coding sequence, overlapping regions allowed alignment of the three clones (Figure 1A-1B) and generation of a putative composite complete coding sequence that encoded for a protein of approximately 122,000 MW. One of these clones (PC5) was used to screen a second cDNA library that had been constructed in a modified lambda ZAP vector, YcDE11. Approximately 1% of the clones hybridized to PC5. Four of these clones GP3, GP22, GP46 and GP14 were sequenced, and all contained open reading frames encoding proteins highly similar to the original gp116 clones (Figure 1A-1D). The clone with the largest insert (GP3), 2869 bases plus a poly-A tail, has an open reading frame encoding for a protein of 104,048 MW. Based on the sequences in the composite protein, GP3 also appeared to be incomplete at the 5' end; PCR was utilized to determine the full sequence of this gene. The 5' end of the message was identified by anchored PCR (24) using primer JK58, which spanned the putative start codon of the composite protein. The intervening region was determined by reverse transcription followed by PCR, using primers spanning the 5' end to base 722 in GP3. A single clone was identified that had an identical sequence to the first 76 bases of GP3. The complete, composite cDNA contained an open reading frame encoding a protein of 122,997 MW.

To demonstrate unambiguously that this cDNA encoded the major surface glycoprotein, fragments of purified gp116 were sequenced (29). Although the amino terminus was blocked, the sequence of an endoproteinase LysC-generated 18 amino acid fragment was obtained. This sequence is identical to amino acids 423 to 440 in the deducted PC5 protein sequence, and is highly conserved in the other six

clones. A second sequenced peptide, identical to residues 365 to 378 of PC5, is much less well conserved among the other clones. An additional peptide (peptide 1), identical to residues 49 to 57 of PC5, was sequenced to show that the protein began with the methionine at position 1, rather than 165 (which would result in a protein of approximately 104,000 MW).

Peptides based on regions in GP3 and PC5 were used to generate antibodies in rabbits. By immunoblot, these anti-peptide antibodies were shown to react with intact as well as lyticase-solubilized gp116 (See Figure 2).

Multiple genes encode the major surface glycoprotein. Antigenic variability of surface proteins is an important mechanism for evading host defenses in a number of organisms (e.g., trypanosome and borrelia species) (31,32). Upon the identification of multiple clones encoding for a family of closely related surface proteins, the present inventors theorized that antigenic variability of surface proteins may be important to P. carinii. Heterogeneity of genes for gp116 may represent the occurrence of multiple alleles for a single-copy gene, multiple genes encoding for a family of related proteins, splicing, or a combination of these factors.

Southern hybridization experiments of P. carinii DNA treated with Nde 1 or Eco R1 and probed with either a conserved 30-mer oligonucleotide (MSG1) or PC5 (blot not shown), revealed multiple bands (Figure 3A), strongly arguing for the presence of multiple genes. When oligonucleotides specific for GP3 or GP14 were utilized in Southern hybridization studies, fewer bands were seen, and the bands were different for the two probes, although all bands detected with these probes were also seen when probing with MSG1. The reactivity of a given band with MSG1 was consistently greater than with the specific oligonucleotides (Figure 3A) despite the fact that all probes had approximately the same specific activity. From these experiments, the inventors concluded that multiple

similar genes encoding the major surface glycoprotein were present in those bands, but the regions corresponding to the specific oligonucleotides were poorly conserved in many of these genes. Hybridization to blots of pulse-field gel-separated P. carinii chromosomes (28) showed the presence of gp116 sequences on multiple chromosomes (Figure 3B), consistent with multiple genes per P. carinii genome. Single copy genes have been previously demonstrated to hybridize to a unique chromosome in all P. carinii isolates (28). Northern hybridization of P. carinii RNA with PC5 revealed a single band of approximately 4,000 bases (Figure 3C), demonstrating that if multiple transcripts are made, they are similar in size. Since P. carinii cannot be grown consistently in vitro, at present it is impossible to clone single organisms to further clarify these issues.

Analysis of the coding regions. The genes encoding gp116 are rich in adenosine and thymidine residues (63% in GP3), and is 69% adenosine or thymidine in the third codon position of GP3, similar to the other coding sequences of P. carinii that have been identified to date (4,5). Nucleotide variability among the clones is not located primarily in the wobble position of the codon: among the 437 nucleotides that differ in PC14 and GP3, for example, 36% are in the first, 30% in the second, and 34% in the third codon position. GP3 has a consensus polyadenylation signal (AATAAA) at nucleotides 3470-3475.

Analysis of the coding sequences shows that of 135 amino acid residues common to all seven clones, only 60 (44%) are identical in all clones, although conservation is higher among pairs of clones. GP46 and GP14 are identical through the first 227 amino acid residues, but subsequently diverge. The cysteine content (5.5%) of the complete protein is relatively high with a very strong conversion of cysteine residues: of 267 residues present in the seven clones, only one is not conserved, and this results from a 16-base frame-shift at nucleotide positions 1045 to 1060 of GP3 compared to PC3 and PC5. The cysteines are

concentrated primarily in three regions: 37 to 243, 329 to 758, and 914 to 941 of GP3. In these regions the cysteine residues do not occur at random, but are most often separated from another cysteine by six (20 occurrences) or seven (six occurrences) amino acids. There is no predictable pattern to the intervening amino acids. The conservation of cysteines, together with their repetitive nature, suggests the occurrence of a repetitive secondary structure, such as loops formed by intramolecular disulfide bonds, that may be functionally important. There is a poorly conserved region rich in proline and glycine residues between residues 817 and 870 of GP3, and a region rich in threonine and serine residues near the carboxy terminus (953-1052 of GP3).

The present inventors and others (8,10,12) have previously shown that gp116 has N-linked carbohydrate residues that account for approximately 10% of its apparent molecular weight. GP3 contains five potential glycosylation sites (Asn-X-Ser/Thr) (Figure 1a). Two of these sites (573 and 809) are conserved in the overlapping regions of the other clones. It is unknown whether O-linked glycosylation sites also exist in gp116. The threonine/serine-rich region may be a candidate for such glycosylation, as has been suggested for a serine-rich region in yeast gp115 and a threonine rich region in the promastigote surface antigen-2 of Leishmania major (33,34).

Analysis of the hydrophilicity profile of the encoded proteins by the algorithm of Kyte and Doolittle (35) demonstrates a single hydrophobic region common to all clones encompassing the last 15 amino acids at the carboxy-terminus. There is no hydrophilic region compatible with an intracytoplasmic domain, nor is there a hydrophobic leader sequence. The position of the hydrophobic tail is consistent with a glycosyl phosphotidylinositol membrane anchorage (36) for this surface protein, although currently there is no evidence to support such a linkage.

Searches of GenBank and PIR failed to identify any significant similarity to other known genes or proteins.

The DNA sequences for the P. carinii major surface glycoprotein, such as shown in Figure 1b, can be modified to provide sequences that are mutants, deletions, or substitutions thereof which encode a protein having at least 90% homology with the naturally occurring major surface glycoprotein and possessing substantially the same properties as the P. carinii major surface glycoprotein.

The major surface glycoproteins of P. carinii preferably comprises one of a homologous variant of said major surface glycoproteins of P. carinii having less than 8 conservative amino acid changes, preferably less than 5 conservative amino acid changes. In this context, "conservative amino aid changes" are substitutions of one amino acid by another amino acid wherein the charge and polarity of the two amino acids are not fundamentally different. Amino acids can be divided into the following four groups: (1) acidic amino acids, (2) neutral polar amino acids, (3) neutral non-polar amino acids and (4) basic amino acids. Conservative amino acid changes can be made by substituting one amino acid within a group by another amino acid within the same group. Representative amino acids within these groups include, but are not limited to, (1) acidic amino acids such as aspartic acid and glutamic acid, (2) neutral polar amino acids such as valine, isoleucine and leucine, (3) neutral nonpolar amino acids such as asparganine and glutamine and (4) basic amino acids such as lysine, arginine and histidine.

In addition to the above mentioned substitutions, the major surface glycoproteins of P. carinii of the present invention may comprise the above mentioned specific amino acid sequences and additional sequences at the N-terminal end, C-terminal end or in the middle thereof. The "gene" or nucleotide sequence may have similar substitutions which allow it to code for the corresponding major surface glycoproteins of P. carinii. Individual base pair changes

or deletions or insertion of the DNA encoding for the major glycoproteins of P. carinii can be made by the methods of site-directed mutagenesis which are well known in the art. See Sambrook et al (39).

In processes for the synthesis of the major surface glycoproteins of P. carinii, DNA which encodes the major surface glycoproteins of P. carinii is ligated into a replicable (reproducible) vector, the vector is used to transform host cells, and the affector is recovered from the culture. The host cells for the above-described vectors include gram-negative bacteria such as *E. coli*, gram-positive bacteria, yeast and mammalian cells. Suitable replicable vectors will be selected depending upon the particular host cell chosen.

SIGNIFICANCE OF EXPERIMENTAL RESULTS

Although P. carinii has been a major pathogen in human immunodeficiency virus of infected patients since the beginning of the AIDS epidemic, inability to culture the organism has made studies of immunopathogenesis very difficult. Experiments investigating host-organism interactions have recently focused on the major surface glycoprotein. Although the function of this protein is unknown, it is an abundant surface-exposed glycoprotein that has the potential to interact with multiple host cell-associated or secreted proteins. As a surface protein, it is likely a primary target of the immune response. The present inventors have shown in the current experiments that multiple genes encode a family of related major surface glycoproteins, and that, based on chromosomal blots, multiple copies of these genes are present in the P. carinii genome. Based on the presence of multiple genes, the present inventors believe that antigenic variability may play a role in immune evasion. Although antigenic variability is well-known in protozoal and bacterial pathogens (31,32), the variability of the major surface

glycoprotein is the first description of this phenomenon in the fungi.

Previous experiments have shown that the major surface glycoprotein of P. carinii obtained from different species vary in size and are antigenically distinct (8,13). However, no experiment has previously suggested variability in the protein moiety of the major surface glycoprotein in organisms obtained from a single species. An epitope of the major surface glycoprotein with a critical carbohydrate component that is conserved in P. carinii isolated from multiple species was identified by monoclonal antibody studies (13), and administration of a monoclonal antibody to this epitope resulted in a decrease in the intensity of infection in two host species (17).

CLONING OF HUMAN ANTIGENS

Based on the above, the corresponding human antigen can be prepared as follows:

Materials

P. carinii organisms could be obtained from human HIV-infected patients and partially purified by Ficoll-Hypaque density gradient centrifugation as described (18).

Human **P. carinii libraries** could be constructed in the same manner as the P. carinii cDNA library in λ ZAP that has been described (4). A second library can be constructed in a similar fashion using oligo-dT-selected mRNA and subcloning into a modified λ ZAP vector (19), YcDE11, which contained sequences necessary for Saccharomyces cerevisiae replication and expression.

General Methods

Several methods could be used to screen the human P. carinii library.

1. The library could be screened with the already identified rat P. carinii surface antigen genes. This could identify the genes since antibody studies have demonstrated that although the rat and human P. carinii

proteins are antigenically different, there is also cross-reactivity, and thus, there is likely to be conservation at the DNA level as well. Once one human P. carinii major surface glycoprotein gene is identified, that gene may be used to identify other members of the gene family.

2. The library may be screened using a conserved oligonucleotide whose sequence is based on the available rat P. carinii major surface glycoprotein genes. Since conserved regions are presumably functionally important, and since the rat and human P. carinii major surface glycoprotein are homologous, they would have conserved the same regions that were conserved among rat P. carinii genes. Low stringency conditions may be used to obtain hybridization even if the conservation is not absolute.

3. Conserved oligonucleotides may be used based on sequences of the rat P. carinii major surface glycoprotein genes as primers for the polymerase chain reaction to be performed using human P. carinii DNA extracts as template. Conditions may be adjusted to low stringency if needed. Once a human P. carinii-specific piece of DNA is amplified, that DNA fragment may then be used to screen the library to identify larger fragments or the entire gene.

4. Amino-acid sequence information from the purified human P. carinii major surface glycoprotein may be obtained by direct sequencing of proteolytic-enzyme generated fragments, in a manner similar to that done with the rat P. carinii major surface glycoprotein. This information may then be used to generate oligonucleotides that may be used either directly to screen the library, or as primers for PCR to amplify a fragment of the human P. carinii major surface glycoprotein gene, which may then be used for further screening.

The identification of a multi-gene family of proteins is difficult because P. carinii cannot be cultured or cloned. The number of genes per genome encoding the major surface glycoprotein is difficult to estimate based on current data, but Southern blot experiments conducted by

the inventors using both conserved and specific oligonucleotides have led the inventors to believe that many similar genes exist in organisms obtained from a single host. The use of antibodies raised against peptides or oligonucleotides specific for individual genes will help determine if single organisms are expressing one or more genes, or if expression of specific genes is associated with specific stages of P. carinii.

P. carinii has been one of the most devastating complications of the immunosuppression associated with human immunodeficiency virus infection. The use of chemoprophylaxis has lead to a marked decline in the incidence of P. carinii pneumonia, but the agents used for prophylaxis are associated with significant adverse reactions or a high failure rate (37). The recent demonstration that novel, potentially protective, immune responses to HIV can be induced by immunization of HIV-infected patients with rgp160 (38) suggests that immunoprophylaxis may also be an effective alternative for controlling HIV-related opportunistic infections. The major surface glycoprotein of P. carinii can be used as a vaccine and as a diagnostic reagent. Additionally, the detailed study of this protein and its expression should lead to an understanding of its functional role in the pathogenesis of P. carinii pneumonia, and may lead to novel strategies designed to prevent or control P. carinii infection and its devastating consequences.

In use as a vaccine, the P. carinii major surface glycoprotein antigen of this invention can be administered to mammals; e.g., human, in a variety of ways. Exemplary methods include parenteral (subcutaneous) administration given with a nontoxic adjuvant, such as an alum precipitate or peroral administration given after reduction or ablation of gastric activity; or in a pharmaceutical form that protects the antigen against inactivation by gastric juice (e.g., a protective capsule or microsphere).

The dose and dosage regimen will depend mainly upon whether the antigen is being administered for therapeutic or prophylactic purposes, the patient, and the patient's history. The total pharmaceutically effective amount of antigen administered per dose will typically be in the range of about 5μg to 1280μg per patient.

For parental administration, the antigen will generally be formulated in a unit dosage injectable form (solution, suspension, emulsion) in association with a pharmaceutically acceptable parenteral vehicle. Such vehicles are inherently nontoxic and nontherapeutic. Examples of such vehicles include water, saline, Ringer's solution, dextrose solution, and 5% human serum albumin. Non-aqueous vehicles, such as fixed oils and ethyl oleate, may also be used. Liposomes may be used as vehicles. The vehicle may contain minor amounts of additives, such as substances which enhance isotonicity and chemical stability; e.g., buffers and preservatives.

The recombinant major surface glycoprotein of this invention can provide a reagent to be used in a variety of diagnostic assays to detect antibodies to P. carinii as well as being useful in developing additional reagents that can detect antigens in clinical specimens. The recombinant protein, can be used directly in assays to detect anti-P. carinii antibodies. Such assays would include, for example, ELISA (enzyme-linked immunosorbent assays), western blot (immunoblot) and immunoprecipitation assays. For antigen detection, antibodies, either polyclonal or monoclonal antibodies, can be generated to the recombinant proteins. These antibodies can then be used in antigen-capture assays using, for example, an ELISA format, and in immunofluorescent assays.

The sequences of the genes can also be used to make primers for use in polymerase chain reaction studies for the diagnosis of P. carinii infection as well as to make oligonucleotide probes that can be used directly in diagnostic assays for detecting the DNA of P. carinii.

The invention being thus described, it will be obvious that the same may be varied in many ways. Such variations are not to be regarded as a departure from the spirit and scope of the invention, and all such modifications as would be obvious to one skilled in the art are intended to be included within the scope of the following claims.

LITERATURE CITED

1. Gottlieb, M.S., Schroff, R., Schanker, H.M., Weisman, J.D., Fan, P.T., Wolf, R.A. & Saxon, A. (1981) *N. Engl. J. Med.* **305**, 1425-1431.

2. Masur, H., Michelis, M.A., Greene, J.B., Onorato, I., Stouwe, R.A., Holzman, R.S., Wormser, G., Brettman, L., Lange, M., Murray, H.W. & Cunningham Rundles, S. (1981) *N. Engl. J. Med.* **305**, 1431-1438.

3. Masur, H., Lane, H.C., Kovacs, J.A., Allegra, C.J. & Edman, J.C. (1989) *Ann. Intern. Med.* **111**, 813-826.

4. Edman, J.C., Edman, U., Cao, M., Lundgren, B., Kovacs, J.A. & Santi, D.V. (1989) *Proc. Natl. Acad. Sci. USA.* **86**, 8625-8629.

5. Edman, U., Edman, J.C., Lundgren, B. & Santi, D.V. (1989) *Proc. Natl. Acad. Sci. USA.* **86**, 6503-6507.

6. Edman, J.C., Kovacs, J.A., Masur, H., Santi, D.V., Elwood, H.J. & Sogin, M.L. (1988) *Nature* **334**, 519-522.

7. Stringer, S.L., Stringer, J.R., Blase, M.A., Walzer, P.D. & Cushion, M.T. (1989) *Exp. Parasitol.* **68**, 450-461.

8. Lundgren, B., Lipschik, G.Y. & Kovacs, J.A. (1991) *J. Clin. Invest.* **87**, 163-170.

9. Gigliotti, F., Ballou, L.R., Hughes, W.T. & Mosley, B.D. (1988) *J. Infect. Dis.* **158**, 848-854.

10. Tanabe, K., Takasaki, S., Watanabe, J.I., Kobata, A., Egawa, K. & Nakamura, Y. (1989) *Infect. Immun.* **57**, 1363-1368.

11. Linke, M.J., Cushion, M.T. & Walzer, P.D. (1989) *Infect. Immun.* **57**, 1547-1555.

12. Radding, J.A., Armstrong, M.Y.K., Ullu, E. & Richards, F.F. (1989) *Infect. Immun.* **57**, 2149-2157.

13. Gigliotti, F. (1992) *J. Infect. Dis.* **165**, 329-336.

14. Pottratz, S.T., Paulsrud, J., Smith, J.S. & Martin, W.J.,II (1991) *J. Clin. Invest.* **88**, 403-407.

15. Ezekowitz, R.A.B., Williams, D.J., Koziel, H., Armstrong, M.Y.K., Warner, A., Richards, F.F. & Rose, R.M. (1991) *Nature* **351**, 155-158.

16. Zimmerman, P.E., Voelker, D.R., McCormack, F.X., Paulsrud, J.R. & Martin, W.J.,II (1992) *J. Clin. Invest.* **89**, 143-149.

17. Gigliotti, F. & Hughes, W.T. (1988) *J. Clin. Invest.* **81**, 1666-1668.

18. Kovacs, J.A., Halpern, J.L., Swan, J.C., Moss, J., Parrillo, J.E. & Masur, H. (1988) *J. Immunol.* **140**, 2023-2031.

19. Short, J.M., Fernandez, J.M., Sorge, J.A. & Huse, W.D. (1988) *Nucleic Acids Res.* **16**, 7583-7600.

20. Young, R.A. & Davis, R.W. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80**, 1194-1198.

21. Beall, J.A. & Mitchell, G.F. (1986) *J. Immunol. Methods* **86**, 217-223.

22. Feinberg, A.P. & Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6-13.

23. Sanger, F., Nicklen, S. & Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463-5467.

24. Kawasaki, E. S. (1990) in *PCR protocols: A guide to methods and applications, eds. Innis,* M.A., Gelfand, D.H., Sninsky, J.J. & White, T.J. (Academic Press, Inc., San Diego, CA), pp. 21-27.

25. Loh, E.Y., Elliott, J.F., Cwirla, S., Lanier, L.L. & Davis, M.M. (1989) *Science* **243**, 217-220.

26. Davis, L.G., Dibne, M.D. & Battey, J.F. (1986) *Basic Methods in Molecular Biology, Elsevier,* New York.

27. Volpe, F., Dyer, M., Scaife, J.G., Darby, G., Stammers, D.K., and Delves, C.J. (1992) *Gene* **112**, 213-218.

28. Lundgren, B. Cotton, R., Lundgren, J.D., Edman, J.C. & Kovacs, J.A. (1990) *Infect. Immun.* **58**, 1705-1710.

29. Lane, W.S., Galat, A., Harding, M.W. & Schreiber, S.L. (1991) *J. Prot. Chem.* **10**, 151-160.

30. Staros, J.V., Wright, R.W., and Swingle, D.M. (1986) *Anal. Biochem.* **156**, 220-222.

31. Boothroyd, J.C. (1985) *Ann. Rev. Microbiol.* **39**, 475-502.

32. Barbour, A.G., Barrera, O. & Judd, R. (1983) *J. Exp. Med.* **2127**, 2140.

33. Vai, M., Gatti, E., Lacana, E., Popolo, L. & Alberghina, L. (1991) *J. Biol. Chem.* **266**, 12242-12248.

34. Murray, P.J. & Spithill, T.W. (1991) *J. Biol. Chem.* **266**, 24477-24484.

35. Kyte, J. & Doolittle, R.F. 1982) *J. Mol. Biol.* **157**, 105-132.

36. Low, M.G. (1987) *Biochem. J.* **244**, 1-13.

37. Kovacs, J.A. & Masur, H. (1992) *Clin. Infect. Dis.* **In press**,

38. Redfield, R.R., Birx, D.L., Ketter, N., Tramont, E., Polonis, V., Davis, C., Brundage, J.F., Smith, G., Johnson, S., Fowler, A., Wierzba, T., Shafferman, A., Volvovitz, F., Oster, C. & Burke, D.S. (1991) *N. Engl. J. Med.* **324**, 1677-1684.

39. Sambrook, J., Fritsch, E.F., Maniatis, T. (1989) Molecular Cloning, A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory Press, Chapter 15.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Kovacs, Joseph A.
Angus, C. W.
Powell, Francoise
Edman, Jeffrey C.

(ii) TITLE OF INVENTION: GENES THAT ENCODE A SURFACE PROTEIN OF P. CARNII

(iii) NUMBER OF SEQUENCES: 19

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: Birch, Stewart, Kolasch & Birch
(B) STREET: 8110 Gatehouse Road
(C) CITY: Falls Church
(D) STATE: Virginia
(E) COUNTRY: USA
(F) ZIP: 22042

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.25

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER: US 07/958,683
(B) FILING DATE: 09-OCT-1992
(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Murphy Jr., Gerald M.
(B) REGISTRATION NUMBER: 28,977
(C) REFERENCE/DOCKET NUMBER: 1173-368P

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 703-205-8000
(B) TELEFAX: 703-205-8050

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2110 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(iii) HYPOTHETICAL: NO

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Pneumocystis carinii

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

TGGTTATCCT CGTGGAGGTC ATTTGATGAA ATATATGAAG GAGGCGATAT AAGTTTTGAT 60

CATGAAAAAC TCGAATTTAA CGAATATAAT CAAGTTTTAC AAATGCTTGA AAAGGCAAAA 120

AATTGGGAAC CGGCTTTGTT GATAGAACCA AAGATTTTTC TAATAGACGA TATGAAGGGA 180

GAATTGAGTT AAATCATTTG GGGAGACGCC CAGGAGTCGA CTATTTTAGG AAAGGTGGGG 24

ATGTTTTTAC TGATGGTTAT CCTCGTGGAG GTCATTTGAT CGAGGATGAG TTGTCCGAAG 30

AGGCGGCAAT GGCACGGCCG GTTAAGAGGC AAGCAAAAGT AGTACAAGGA GCACAAGGAG 36

CACAAGATGA CATTAAGGAG GAACACCTTT TGGCTTTCAT TGCGAAGAAG GAATATAGTA 42

ATGAGGATAA ATGCAAACAA GAACTCAAGA AATATTGTGA AGAGTTGAAG GAAGCAGATG 48(

GTAAATTCAA TGTTAATGAT AAAGTTAAAG AACTTTGTGG TGGTGGTGAT GAAGCAAAAC 54(

GAGATAAAAA ATGCAAAGAC CTGAAAGACA AAGTTGAAGA TGAATTAGAA AATTTTGATG 60(

ATGAACTTCA AGAAGCATTG AAAGACATAA AAGATGAAAA TTGTGAAAAA CATGAAGAAA 66(

AATGTATACT TTTAGAAGAC ACGGGTTATA GTGAAGATAT TAAGAAGAAC TGTGTCAAGT 720

TGAGGGAAGG ATGTTACAAA TTGAAGCGTA AAAAGGTGGC AGAGGAGCTC CTTTTGAGGG 780

CGCTCGGAGG GGATGCTAAA GATGAAGCTA AATGTAAAGA AAAGATGAAA ACTGTTTGCC 840

CAATGTTAAG CCGAGAAAGT GACGAGCTGA TGTTTTTCTG CCTTGATTCG GATGGAACGT 900

GTAAAGCGCT GAAAACAAAA TCAGAAGAAG TTTGCCTGCC TTTAAAAGAA AAGCTTAAAG 960

ATGGCGAATT AAAGGAAAAA TGTCATGAAA GACTTGAGAA ATGTCATTTT TACAAAGAAG 1020

CGTGTACTGA AACAAAGTGT GATGAGGATA TGAAGCAATG CAAGGAAAAA GGATTCACAT 1080

ATAAAGCGCC GGAATCTGAT TTTAGTCCTG TCAAGCCGAA GGCGTCGTTG TTGAGAAGTA 1140

TTGGGTTGGA TGATGTGTAT AAAAAGGCTG AAAAAGAAGG AATTATTATT GGAAAATCAG 1200

GAGTGGATCT ACCAAGGAAG TCAGGTACAA AATTTCTGCA AGATCTCTTG CTACTGTTGA 1260

GCAGAGATGA GAATGATGCA GGGAAGAAAT GCGGTAAAGC GTTAGGAAAA TGTGAAACTT 1320

CTAAGTATTT GAATACTGAT TTGATGGAGT TATGCAAAGA TGCTGATAAA GAAAATAAAT 1380

GCAAAAAAAA GCTAGATGTA AAAGAAAGAT GTACAAAACT CAAGTTAAAT CTTTATGTGA 1440

AAGGGTTGTC TACGGAGTTT AAAGAAGATA AAAAATCACA TCTTTTATCG TGGGGACAGC 1500

TTCCAACATT ATTTACGAAG GGAGAGTGTG CAGAACTTGA GTCGGAATGT TTCTATTTAG 1560

AAAATGCGTG TAAAGATAAT GAGATTGGTG AAGCGTGTCA AAATCTACGA TCAGCGTGCT 1620

ATAAAAAGGG ACAAGACAGG ATGTTGAATA AGTTCTTTCA AAAGGAATTG AAGGGAAAGC 1680

TTGGTCATGT AAGATTTTAT AGCGATCCTA AAGATTGTAA AAAATATGTG GTAGAAAACT 1740

GTACAAAACT TAAAAAAGAT AAAAGATACC TTTCAAAATG TCTTTATCCT AAAGAACTAT 1800

GTTATGGGCT TTCAAATGAT ATTTTTCTCC AATCCAAAGA GTTAAGTTCG CTTTTAGATG 1860

ATCAGAGAGA TTTTCCATTT GAAAAGGATT GTCTTGAATT GGGAGAGAAG TGTGATCAAC 1920

TTAGTAGTGA TTCATTATTG AATTTAGAAA AGTGTATAAC ATTGAAAAGA CGCTGTGAAT 1980

ATTTTGACGT TACAGAAAGA TTTAGAAAAG TATTTTTAGA AAAAAAGGAT GATTCGTTAA 2040

TGATTCAGGA AAACTGTACA AAGGCATTGC ATGAGAAATG TAATACTTTA TATAAGAGGA 2100

GAAAGAATTC 2110

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1454 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(iii) HYPOTHETICAL: NO

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Pneumocystis carinii

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

GAAGAAGAAT TGACAACTTT TGAAGGGGAT CTTGACACTG CATTGAAAAA TGGCATAAAA 60

GATGAAGATT GTGAAAAACA TGAAGAAAAA TGTATACTTT TAGAGGAAGC AGACCCAAAT 120

AGTCTTAAGG AGAAGTGTGT CAAGTTGAGG GAAGGATGTT ACGAATTGAA GCGTGAAAAG 180

GTGGCAGAGG AGCTCCTTTT CAGGGCACTC GGAGGGGATG CTAAAGAAGA TGGTAAATGT 240

AAAGGAAAGA TGAATACTGT TTGCCCAGTG TTAAGCCGAG AAAGCGACGA ATTGATGACT 300

TTTTGCCTTG ATCCGGATGG AACGTGTGGA GAGCTGAAAA CAAAATTGGG CGAAGTTTGC 360

AAACCTTTAG AAACAGAATT GAATGAGAAA AGCTCAGAAA AGTGTCATGA AAGACTTGAG 420

AAATGTCATT TTTACAAAGA AGCGTGTGGT AATACAAAAT GTAAGGAGGA TAAGACGAAA 480

TGCGAGGAAA AAGGATTCAC ATATAAAGCG CCGGAATCTG ATTTTAGTCC GGTCAAGCCG 540

AAGGCGTCGT TGTTGAGAAG TATTGGGTTG GAAGATGTGT ATAAAAACGC GGAAAAACAT 60(

GGGATTATTA TTGGAAAATC AGGAGTGGAT CTACCAAGGA AGTCAGGTAC AAAATTTCTG 66(

CAAGATCTCT TGCTAGTCTT GAGCAGAGAT GAGAATGATG CAGGGAAGAA ATGCGGTAAA 72

GCGTTAGGAA AATGTGATGC TTCTAAGTAT TTGGATCATA ATTTGAAAGA GTTATGCAAT 78

GATGGAAAGA AAAACGACAA ATGCAAAGAA TTACTAGATG TAAATGTAAA AGAAAGATGT 84

ACAAAACTCA AATTAAATCT TTATGTGAAA GGGTTGTCTA CAAAATTTGA AAAAGCTGAA 9(

AAATCAGATC TTTTATCGTG GGGACAGCTT CCAACATTAT TTACGAAGGG AGAGTGTGCA 9(

GAACTTGAGT CGGAATGTTT CTATTTAGAA AATGCGTGTA AGGATAATAA GATTGATGAA 10:

GCATGTCAAA ATGCAAGAGC AGCGTGCTAT AAAAAGGGAC AAGACAGGAT GTTGAATAAG 10

TTCTTTCAAA AGGAATTGAG GGGAAATCTT GGTCTTGTAA GATTTTATAG CGATCCTGAA 11

GAATGCAAAA AATCTGTGGT AGGAAACTGT ACAAAACTTA AAGAAGATAG TAGATACCTT 12

TCAAAATGTC TTTATCCTAA AGAATTATGT TATGCGCTTT CAAATGATAT TTTTCTTCAA 12

TCCAAAGAGT TAAGTTCGCT TTTGGATGAT CAAAGGGATT TTCCATTAGA AAAGGATTGT 12

CTTGAATTGG TGGAGAAGTG TGATGAACTT AGTAGTGATT CATTATTGAA TTTAGAAAAG 13

TGTATAACAT TGAAAAGACG CTGTGAATAC TTTAAGGTTA CAGAGGGATT TAGAAAAGTA 1·

TTTTTAAAAA AAAA 1·

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2190 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(iii) HYPOTHETICAL: NO

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Pneumocystis carnii

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

TTCCCAAAAC TTTTATCGTG GGGACAGCTT CCAACCCTTT TTATAAAAGG AGAGTGTGCA 60

GAACTTGAGT CGGAATGCTT CTATTTAGAA AATGCGTGTA CGAATAAGAT TGGTGAAGCA 120

TGTCAAAATG TACGATCAGC GTGCTATAAA AAGGGACAAG ACAGGATGTT GAATACGTTG 180

TTTCGAGAGG AGATGAAGGG AAAGCTTGCT AATATAAAAT ATTTTAATGA TACTGAAAGT 240

TGCAAAAAAT CAGTGGCAAA AAAGTGTGCA GAACTTGATA AAAGATACCT TTCAAAATGT 300

CTTTACCCTA AAAGACTATG TTATGTGCTT TCAGATGATA TTTTTCTTCA ATCAAAAGAG 360

TTAAGTGTGC TTTTAGATGA TCAGAGAGAT TTTCCATTAG AAAAGGATTG TGTTGAATTG 420

GGAGAGAAGT GTGATGAACT TGGTAGTGAT TCATTATTGA ATTTAGAAAA GTGTATAACA 480

TTGAAAAGAC GCTGTGAATA CTTTAAGGTT ACAGAAAGAT TTAGAAAAGT ATTTTTAGAA 540

AGAAAAGATC ATTCATTATA CGATGAGCAA AATTGTACGA AGGCGTTGCA TGAGAAATGT 600

GAAGCTTTAT TTAGGAAAAG GAGGAATCCA TTTGAGTTTT CATGTGCTTT GCAAGAAGAA 660

ACATGTCAAC GTATGGTATA CCATACAACT CAAGATTGTA TTTATTTAAA AGACAACATC 720

AAAAATAAAA AAATTCTAGA ACAAATTGGA AAAGTAAAAC AGGATAAATC AAAAGAAGCA 780

GAAGTAGAAG AACTCTGCAC AACATGGGGT AAATATTGTC ACCAACTTAT GGAGAATTGT 840

CCAGATAAGT TGAAAAAAAA AAAAAAAAAA GACAATGACA ATAATCAAAA CTGCGAAGAA 900

CTCGAAAAAA AATGCACTGA TACCTTTAAA AAGTTGGAAT TGAAGGATGA GCTGACTCAT 960

CTGTTGAAAG GCAGCTTAAA GGATAAAGAA AAATGTAAAG TAACACTAGG ACAGCGTTGC 1020

CCTGAGTTGA AAAATAATGA TACATTCAAA ATTCTGCTTA CTAATTGTGA AGATTCCTTG 1080

GAAAATGTTT GCGCGGAATT AGTTAAAAAA GTACAGAAGA AATGTCCTAC TTTAAAAGAC 1140

GAACTGAATA AAGCGAAAGA TGAGTTGACA AAGATGAAGA CTGAGTACGA AAATGCTAAA 1200

AAGGCGGCAG AAGAATCTAC AAACAAAGCT AGCTTATTGC TATCAAAGTC TGGAAAAGCC 1260

GCAATGCCAA CTGCGCAGAA TGGCAGTGCT TCTGCACCAC CATCAGCACC AGCAGAATCA 1320

GGATCATCAC CAGCATCAGG GTCACCACCA GCATCAGAGC CATCAACTAA TGGAAAGGTG 1380

GACACGCCAG CTGGAGGATC AGGGACACAA GATAAAACAT CAGACGCATC AGGTCAAACG 1440

ACGAAGTATA CAAAACTTGG ACTCGTTAAA AGAGCATATG TAGCTGAAGG AGTATCAGAA 1500

GAAGAGGTAA AAGCATTTGA TGCAACGACG GTAGCATTGG AATTGTATTT GGAATTGAAA 1560

GAGGAATGCA ATGCTTTAGA ACTAGATTGC GGTTTTAAAG AGGATTGTGA GGAATCTAAA 1620

CCAGCTTGTA AAGAAATAGA AGAGTTATGC AAAGGAATAG AATCATTAAA AGTTACGCCT 1680

CATCATACAG AGACGCAAAA AGAAATCTCA ACCACTACGA CGACCACTAC TACGACCACT 1740

ACCACGACTA CTACCACGAC TACGACGACA ACTACTACTA CAACCAAGCC GGGAAGTGGA 1800

GGAAAAGTAA CAGAAGAGTG TACAATGATA CAAACAACAG ATACATGGGT GACACGTACG 1860

TCATTGCATA CGAGTACGAC AACGAGTACG TCGACAGTGA CGTCGACAGT GACATTGACG 1920

TCGATGCGCA AGTGCAAGCC TACCAAATGT ACCACTGATT CAAGCAAAGA GACAGAAGAA 1980

GGAGGAAAAG AAGAAGAAGA AGTAAAACCG AATGATGGGA TGAAAATAAG AGTTCCTGAT 2040

ATGATTAAAA TAATATTGTT GGGAGTGATT GTTATGGGGA TGATGTAAAA TGAATGAAAA 2100

AAATGTTAAT AGAATGAAAA TGTGCATATA TCCATTGTTT ATATATAATA GAAATCTAAA 2160

TGAATGAAAT GAAGTTTTAA TAATTTTAAG 2190

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 3521 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(iii) HYPOTHETICAL: NO

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Pneumocystis carnii

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

TAGACGATAT GAAGAGAGAA TTGAGTTAAA TCATTTGGGG AGACGCCCAG GAGTCGACTA 60

TTTTAGGAAA GGTGGGGATG TTTTTACTGA TGGTTATCCT CGTGGAGGTC ATTTGATCGA 120

GGATGAGTTG TCCGAAGAGG TGGCAATGGC ACGGCCGGTT AAGAGGCAAG CAGTACAAGG 180

AGCACAAGAT GAGATTGATG AGAAACACCT TTTGGCTTTC ATTGTGAAGG ACAAATATAA 240

AGAAGAACAA AAATGCAAAG AAGAACTCGA GAAATATTGT AAAGAGTTGA AGGAAGCAGA 300

TAAAAATCTA GAGAATGTGG ATGATAAAGT TAAAGGGCTT TGTGATGATA AAAAACGAGA 360

CGAAAAATGC AAAGACGTGA AAAAAAAAGT TGAAGATGAA TTAAAAGATT TTGAAGAGGA 420

ACTTCAAAAA GTATTGAATA ATATAAAAGA TGAAAATTGC GAAAAATATG AAGAAAAATG 480

TATACTTTTA GAAGAGACGG ATTATGATGT TATTAAGGAT AACTGTATCG AGTTGAGGGA 540

AGGATGTTAC AAATTGAAGC GTGAAAAGGT GGCAGAGGAG CTTCTTCTGA GGGCGCTCGG 600

AGGGGATGCT AAAGAAGAAG CTAAATGTAA AGGAAAGATG AATACTGTTT GCCCAGTGTT 660

GAGCCGAGAA AGCGACGAAT TGATGTCTTT TTGCCTTGAT TCTGCTAAAA CATGTGGAGA 720

TCTGAAAAAA AAATTGGGTA CTGTTTGCGA GCCTTTAAAA AAAGAGCTTA AAGATAACGA 780

ATTAGCGGAA AAGTGTCATG AAAGACTTGA GAAATGTCAT TTTTACGGAG AAGCGTGTGA 840

TGATGCGAAA TGCAAGAAGT TTGAGGAGCA ATGCAAGGGA AAAAATATTA TATATAAAGC 900

GCCAGAATCT GATCTTAGTC CTGTCAAGCC GAGGGCGTCC TTGTTGAGAA GTATTGGGTT 960

GGATGATGTG TATAAAAACG CGGAAAAACA TGGGATTATT ATTGGAAAAT CAGGAGTGGA 1020

TCTACCAAGG AAGTCAGGTA CAAATTTCTG CAAGATCTCT TTGCTACTGT TGAGCAGAGA 1080

TGAGGATAAG AAGGAACCAG ATAAAAAGTG CACTAAAGCG TTAGAAAAAT GTGATGCCTC 1140

TAAGTATTTG AATACTGAAT TGGAAAAGTT ATGTAAAGAT GGAAACAAAA ACGAAAAATG 1200

CAAAAAAATA TTAGATGTAA AAGAAAGATG TACAAATCTC AAATTAAAAC TTTATCTGAA 1260

AGGATTGTCT ACGGAATATG ATGATCAAGA ATCAGATCCT TTATCGTGGG GACAGCTGCC 1320

AACTTTTTTT ATAAAAGGAG AGTGTGCAGA ACTTGAGTCG GAATGTTTCT ATTTAGAAAA 1380

GGCGTGTAAA GATAATAATA TTGATAAAGC GTGCCAAAAT GCAAGAGCAG CGTGCTATAA 1440

AAAGGGACAA GACAGGATGT TGAATAAGTT CTTTCAAAAG GAATTGAAGG GAAAGCTTGG 1500

TCATGTAAGA TTTTATAGCG ATCCTAAAGA TTGTAAAAAA TATGTGGTAG AAAACTGTAC 1560

AAAACTTGAT AAAAAATATC TTCCACGATG TCTTTATCCT AAAGAACTAT GTTATGGGCT 1620

TTCAAATGAT ATTTTTCTTC AATCCAAAGA GTTAAGTGCG CTTTTGGATG ATCAAAGGGA 1680

TTTTCCATTA AAAAAGGATT GTGTTGAGTT GAAGGAGAAG TGTGATGAAC TTAGTAGTGA 1740

TTCATTATTG AATTTAGAAA AGTGTATAAC ATTGAAAAGA CGTTGTGAAT ACTTTAGAGT 1800

TTCAGAGGGA TTTAGAAATG TATTTTTAGA AAAAAAGGAT GATTCGTTAA TGACTCAGGA 1860

TAACTGTACA AAGGCATTGC ATGAGAAATG CCATCAATTA TATAGGAGGA GAAAGAATTC 1920

ATTTAGTGTT TCATGTGCTT TACCAGAAGA AACATGTAGT TATATGGTAT TCCATACAAG 1980

TCAAGATTGT AGTAGTTTAA AAGTCAACAT CAAGAATGAA AAAATTCTAG AAAAAATTGG 2040

AGAAGAAATT AAAAAAGCAA ATAAAAATGA AGCCTTGGTT GAAGAACTCT GCACAACATG 2100

GGGCCGACAT TGTCACCAAC TTATGGAGAA TTGTCCGGAT GACTTGAAAA AAAAAGAGAA 2160

TGGCAATGGC AATGATCATA ACTGCGAAGC ACTCCAAGAA AAATGCAATA AAACCTTTGA 2220

AAAGTTGAAA TTAGAGGAGG AGCTGAGTCA TCTGTTGAAA GGCAGTTTAA AGGATGATAA 2280

ATGTAAAGAA GCATTAGGAA AGCGTTGCAC TGAGTTGGAA AAGAATGAAG CATTCAAAAC 2340

TCTGTATGGT AAATGTGATG ATAATACCAA GGAAAATGTT TGCAAAAAAT TAGTTGATAA 2400

AGTAAAAAAG AGATGCCCTA CTTTAAAAGA CGAACTGGAG AATGCGAAAA AAGAGTTGAC 2460

AAAGATGAAG AATGAGTACG ATGATCTCAA AAAGGCGGCA GAAAAATCTA CGGAGGCAGC 2520

TAAGTTATTG CTATCAAGAC CTAGACAAAC TGTAATGCCA AATGCGCAGA ATGGCAGTGA 2580

TTCTACACTA GTACCACCAC CACCACAAGC ACCAGCAGGG CCACCACCAC CAGGGTCACC 2640

ACCACCACCA CCATCACAAA ATGGAACGCC AGGCACACCA GGTGGAGAAA CAGGCGCATC 2700

AGGTGGAACA CCAGGCACAC CAGGCACACC AGGCACACCA GGCACACCAG GTGGAATGAT 2760

GAAGTATGCA AAACTTGGAC TCGTTAAAAG AACGTATGTA GATGGAGGTG TATCAGAAGT 2820

AGAGGTCAAA GCATTTGATG CAACGACGAT AGCATTGGAA TTGTATTTGG AATTGAAAGA 2880

AGAATGTAAA GCTTTAGAAT TAGATTGCGG TTTTAAAGAG GATTGTCCAG ATACTAAACA 2940

AGCTTGCGAA AATATAGACA CTTTATGTAA ACTGGAACCA TTAGAAATTA AGCCTCATCA 3000

TACAGAGAAA ATAACAGAAA CAAAGACGGA AACGAAGACG GAAACAAAGA CGGAAACAAA 3060

GACTGATGGC AAGGCTGATG AAAAGACCGT TGAGAAGACT GTTACAGAAA CCAAGTCAGT 3120

AGGTGGAGGA AAAGTAACAG AAGAGTGTAC AATGATACAA ACAACAGATA CATGGGTGAC 3180

GAGTACGTCA TTGCATACGA GTACGACAAC GAGTACGTCA ACGGTGACGT CGACAGTGAC 3240

GTTGACTTCG ATGCGCAAGT GCAAGCCTAC CAAATGTACC ACCGATTCAA GCAAAGAGAC 3300

ACAGAAAGAA GAAGATGATG AAGAAGTGAA ACCGAATGAG GGAATGAAAA TAAGAGTTCC 3360

TGATATGATT AAAATAATGT TGTTGGGAGT GATTGTTATG GGGATGATGT AAATGAATGA 3420

AAAAAATGTT AATAGATTGA AAATGTGCAT ATATCCATTG TTTATATATA ATAAAAATGT 3480

AAATGAATGA AATGAAAAAA AAAAAAAAAA AAAAAAAAAA A 3521

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2058 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(iii) HYPOTHETICAL: NO

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Pneumocystis carnii

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

ATGTTGAGTA CGTTGTTTCG AAAGGAATCG AAGGGGGAGT CTGGTCATAA AAGATATTAT 60

AACCATCCTG AGGAATGCCA AAAATCTGTG GTAAAAGACT GTAAAAAACT TGAAAATAAA 120

GATAAGAGAT ACCTTCCAAA GTGTCTTTAT CCTAAAGAAC TATGTTATAT GCTTTCAGAT 180

GATATTTTCC TTCAATCCAA AGAGTTGGGA GCGCTTTTGG ATGATCAAAG GGATTTTCCA 240

TTAGAAAAGC ATTGTGTTGA ATTGAAGGAG AAGTGTGATG AACTTGAAAC TTATTCACAT 300

TCGAATTCGG AAAAGTGTAT AACATTGAGA AGGCGCTGTG AATACCTTAG AGTTTCAGAG 360

GAATTTAGAA AAGTATTTTT AAAAAGAAAA GATCATGCAT TATATAATGA GCAAAACTGT 420

ACGGAGGTGT TGCAAGAAAA ATGTAATACT TTATATAGGA GGAGAAAGAA TTCATTTAGT 480

GTTTCATGTG CTTTGCCAGG AGAAACATGT GAATATATGG TATACCGTAC AAAAGATGAA 540

TGTTTTTATT TAAGTGGCAA CATGGAGGAT GAAAAAATTG TAGAAGAAAT TGGAAAGAAA 600

AAAGCAAATG AAACAGCACT CGAAGAACTC TGCACAACAT GGGGCCGACA TTGTCACCAA 660

CTTATGGAGA ATTGTCCGGA TGACTTGAAA AAAAAAGAGA ATGGCAATGA CAATGATCAT 720

AACTGTGAAG AACTCGATGA AAAATGCAGT GATACCTTTA AAAGGTTGAA ATTAGAGGAG 780

GAGCTGACTC ATCTGTTGAA AGGCAGCTTA AAGGATAAGG ATGAATGTAA AAAAACATTA 840

GAAAAGCGTT GCACTGAGTT GCAAAATAAT GAAACATTTA AAAATCTGCT TAGTTATTGT 900

GGAGAGAATG ACAAGGGAAC TGTTTGCGAA AAATTAGTTG AAAAACTAAA AAAGAGATGT 960

CCTACTTTAA AAGACGGACT GAATAAAGCG AAAGATGAGT TGACAAAGAT GAAAAAAGAA 1020

TACGATGCGC TTAAAAAGGC GGCAGAAGAA TCTACAAAGG AAGCTAGCTT ATTGCTATCA 1080

AGACCTAGAC AAACTGTAAT GCCAAGTGCG CAGAATGGCA GTGCTTCAGA GCAAGTATTA 1140

CAACCAGTAC AACCAGAATC AGGGTCATCA TCAGGGTCGC CATCATCACC ACCAGGGCCA 1200

CCATCAGCAC CACCACAAAA TGGAACGCCA GCCACACCAG GTGGAGCACC AGGCACACCA 1260

AGCAGTGGAA CGACGGGCCC TGCAAAACTT GGACTCGTTA AAAGAGCATA TGTAGCTGAA 1320

GGAGTATCAG AAGCAGAGGT CAAAGCATTT GATGCAACAA CGATAGCATT GGAGTTGTAT 1380

TTGGAATTGA AAGAAGAATG TAAAGCTTTA GAATTAGATT GCGGTTTTAA AGAGGATTGT 1440

AAGGAAACTG AACCAGCTTG TAAAGAAATA GAAAAGTTAT GTAAACTGGA AGCATTAAAA 1500

GTTGCGCCTC ATCATACAGA GACAATAACA AATAAGGTGA CGGAAACACA GACGGAAACA 1560

AAGACCGTTG AGAAGGTCGA TGACAAGGCT GATGTGAAGA CCGTTGAGAA GACTGTTACG 1620

GTAACCAAAC CAGGAAGTGG AGAAAAAGTA ACAGAAGAGT GTACAATGAT ACAAACAACA 1680

GATACATGGG TGACAAGCAC GTCATTGCAT ACGAGTACGA CAACGAGTAC ATCGACGGTG 1740

ACGTCGACAG TGACGTTGAC CTCGATGCGC AAGTGCAAGC CTACCAAATG TACTACTGAT 1800

TCAAGCAGAG AGACAGATAA AGGAGGAGAA GGAGAAGAAG ATGTAAAACC GAATGAGGGA 1860

ATGAAAATAA GAGTTCCTGA TATGATTAAA ATAATGTTGT TGGGAGTGAT TGTTATGGGA 1920

ATGATGTAAA ATGAATGAAA AAAATGTTAA TAGATTGAAA ATGTGCATAT ATCCATTGTT 1980

TATATATAAT AGAAATCTAA ATGAATGAAA TGAAGTTTTA ATTTTAATAC ACCAAAAAAA 2040

AAAAAAAAAA AAAAAAAA 2058

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2110 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(iii) HYPOTHETICAL: NO

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Pneumocystis carnii

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

GCAGAACTTG TCTCGGAATG TTTTTATTTA GAAAAGGCGT GTAAAGATAA TAAAATTGAT 60

CAAGCGTGTC AAAATGTACG AGCAGCGTGC TATAAAATGG GACAAAATAG GATGTTGAAT 120

ATGCTCTTTC GAGAGGGGTT GAAGGAGAAT TCTGAACGTA TAAAATATTA TGATGAGAAT 180

CCTCAAAAAT GTCAAGAATT TGTGGTAGGA AGCTGTACAA AACTTAAAAA ATATCTTCCA 240

CAATGTCTTT ACCCTAAAGA ACTATGTTAT GCGGTTTCAG ATGATATTTT TCTTCAATCC 300

AAAGAGTTGG GTGTGCTTTT GGATGATCAA AGAGATTTTC CATTAGAAGA GGATTGTCTT 360

GAATTGAAGG AGAAGTGTGC TCAACTTGAA ACTTATTCAA ATTCGAATTC TCAAAAGTGT 420

GCAACATTGA GAAGGCGCTG TAAATACTTA AGAGTTTCTG AGGGATTTAG AAATGTATTT 480

TTAAAAAGAG AAGATGATTC GTTAAAGAAA GAAAACTGTA CGAAGGCATT GCAAGAAAAA 540

TGTGATGCTT TATCTAGGAA AAGGAGGAAT CCATTTGGGT TTTCATGTGC TTTGCGAGAA 600

GAAACATGTG AATATATGGT AGCCCGTACA AAAGACGAAT GTTTTTATTT AAAAGACAAC 660

ATGGAGAATG AAGAAATTCT AAAAGAAATT GAAGAAAAAG CAAAAAAAGA TAATGCAAAT 720

AGAAATGAAA CCTTGGTTGA AGAACTCTGC ACAACATGGG GCCGACATTG TCACCAACTT 780

GTGGGGAATT GTCCGGAGCA GTTGAAAAAA AAAAAAAAAA AAGATGATAA CAAAGATCAT 840

AACTGTGACA AACTCGAAGA AAAATGCAGT GATACCTTTA AAAGGTTGAA ATTAGAGGAG 900

GAGCTGACTC ATCTGTTGAA AGGAAGTTTA AAGAGTGAAG ATGAATGTAA AAAAACATTA 960

GGAGAGCATT GCCCTGAGTT GCAAAAGAAT GATACATTCA AAACTCTGTA TGGTAAATGT 1020

GAAGAGAATG AAAAGGGAAC TGTTTGCAAA AAATTAGTTA AAAAAGTACA AGAGAGATGT 1080

CCTACTTTAA AAACCGATCT GGAGAAGGCG AAAAAAGAGT TGAAGGACAA GAAAGATGAA 1140

TACGATAATG TCAAACAGGC AGCAAAAGAA TCTACGGAGA AAGCTAAGTT ATTACTATCG 1200

AAGCCTCGAC AAACCGTAAC GCCAAATGCG CAGAATGGCA GTGCTTCTGG ACCAGTACCA 1260

GCACCAGCAG CACCTCCAGC AGCACCAGAA GCACCAGCAC AGCCACCACC ACCAGCAGGG 1320

CAACCAAGTG GTGAAACATC AAACGTACCA GGTAAAACGC CAAGCAAAGA AGCTGGAACA 1380

CCAAACACAA CAGATGAAAC GACGAAGAAT CCAAGCCTTG GACTCGTTAA AAGAGCATAT 1440

GTAGAAGGAG GTGTATCAGA AGCAGAGGTA AAAGCATTTG ATGCAACGAC AATAGCATTG 1500

GAGTTGTATT TGGAATTGAA AGAGGAATGC AGCGCTTTAC AACTAGATTG CGGTTTTAGA 1560

AAGGATTGTT CGAGTGTTGA AGGTGTTTGC AAAGAAATAG ACAAGTTATG TGAACTGGAA 1620

CCATTAAAAG TTACGCCTCA TCATACAGAG ACAATAACAA ATAAGGTGAC GGAAACGAAG 1680

ACGGAAACAA AGACAGAAAC AAAGACTGAT GACAAGGCTG ATGAGAAGAC CGGTACGAAA 1740

ACTGTTACAG AAACCAAGAC AATAGGTGGA GGAAAAGTAA CAGAAGAGTG TACAATGGTA 1800

CAAACAACAG ATACATGGAT AACACGTACG TCATTGCATA CGAGTACGAC AACGAGCACG 1860

TCAACGGTGA CGTCGACAGT GACGTTGACC TCGATGCGCA AGTGCAAGCC TACCAAATGT 1920

ACCACTGATT CAACCAAAGA GACACAGAAA GAAGAAGATG ATGAAGAAGT GAAACCGAAT 1980

GAGGGAATGA AAATAAGAGT TCCTGATATG ATTAAAATAA TGTTGTTGGG AGTGATTGTT 2040

ACGGGGATGA TGTAAAATGA ATGAAAAAAA TGTTAATAGA TTGAAAATGT GCATATAAAA 2100

AAAAAAAAAA 2110

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2126 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(iii) HYPOTHETICAL: NO

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Pneumocystis carnii

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

GCAGAACTTG AGTCGGAATG TTTTTATTTA AAAAAGGCGT GTAAAGATAA GGAGATTGAT 60

GAAGCATGTC AAAATGCACG AGCAGCGTGC TATAAAGTGG GAAAAGATAG GATGTTGAGT 120

ACGTTGTTTC GAAAGGAATC GAAGGGGGAG TCTGGTCATA AAAGATATTA TAACCATCCT 180

GAGGAATGCC AAAAATCTGT GGTAAAAGAC TGTAAAAAAC TTGAAAATAA AGATAAGAGA 240

TACCTTCCAA AGTGTCTTTA TCCTAAAGAA CTATGTTATA TGCTTTCAGA TGATATTTTC 300

CTTCAATCCA AAGAGTTGGG AGCGCTTTTG GATGATCAAA GGGATTTTCC ATTAGAAAAG 360

CATTGTGTTG AATTGAAGGA GAAGTGTGAT GAACTTGAAA CTTATTCACA TTCGAATTCG 420

GAAAAGTGTA TAACATTGAG AAGGCGCTGT GAATACCTTA GAGTTTCAGA GGAATTTAGA 480

AAAGTATTTT TAAAAAGAAA AGATCATGCA TTATATAATG AGCAAAACTG TACGGAGGTG 540

TTGCAAGAAA AATGTAATAC TTTATATAGG AGGAGAAAGA ATTCATTTAG TGTTTCATGT 600

GCTTTGCCAG GAGAAACATG TGAATATATG GTATACCGTA CAAAAGATGA ATGTTTTTAT 660

TTAAGTGGCA ACATGGAGGA TGAAAAAATT GTAGAAGAAA TTGGAAAGAA AAAAGCAAAT 720

GAAACAGCAC TCGAAGAACT CTGCACAACA TGGGGCCGAC ATTGTCACCA ACTTATGGAG 780

AATTGTCCAG ATAAGTTGAA AAAAGAAAGT GATAACAGAG ATCATAACTG TGACAAACTC 840

GAAGAAAAAT GCAGTGATAC CTTTAAAAAG TTGAAATTGA AGGAGGAGCT AACTCATCTG 900

TTGAAAGGAA GTTTAAATGA TAAAAAAAAA TGTACAGAAA CATTAGGAAA GAATTGCACT 960

GAGTTGCAAA AGAATGATAC ATTCAAAATT CTGCTTAGTG ATTGTAAAGA TTCCTTGGAA 1020

AATGTTTGCA CAAAATTAGT TGAAAAAGTA CAGAAGAGAT GTCCTGCTTT AAAAACCGAT 1080

CTAGAGGAAG CGAAAAAAGA GTTGAAGGTC AAGAAAGAAG AATATGATGC GCTCAAAAAG 1140

GCAGCAGAAG AATCCAGAAA TAAAGCTAGC TTATTGCTAT CAAGGTCTAA ACAAGCCGTA 1200

ACACCAAGTG GACAGAATGG CAGTGATTCT GTACCAGCAC AGGTACAGCC AGCACCAGCA 1260

GGGCCACCAT CAGCACCAGG GTCGCCATCA TCACCACCAT CACAAAATGG AACGCCAGGT 1320

GCACCAGATG GAACGACAGA CACAGCAGGT GGAACGACGA ATAATGCAAA ACTTGGACTC 1380

GTTAAAAGAG CGTATGTAGA TGAAGGTGTA TCAGAAGCAG AGGTAAAAGC ATTTAATGCA 1440

ACGACAATAG CATTGGAATT GTATTTGGAA TTGAAAGAGG AATGCAGCGC TTTACAACTA 1500

GATTGCGGTT TTAAAGAGGA TTGTCCAGAT ACTAAACAAG CTTGTAAAGA AATAGAAGAG 1560

TTATGTAAAC TGGAAGCATT AAAAGTTGCG CCTCATCATA CAGAGACAAT AACAGAAACG 1620

AAGACAGAAA CGAAGACGGA AACAAAGATG GAAACAAAGA CTGATGACAA GGCTGATGAG 1680

AAGACCGGTA CGAAAACTGT TACAGAAACC AAGACAATAG GTGGAGGAAA AGTAACAGAA 1740

GAGTGTACAT TAGTCAAGAC AACAGATACA TGGGTGACGA GTACGTCATT GCATACGAGT 1800

ACGACAACGA GTACGTCAAC GGTGACGTCT ACAGTGACGT TGACCTCGAT GCGCAAGTGC 1860

AAGCCTACCA AATGTACCAC CGATTCAACC AAAGAGACAC AGAAAGAAGA AGATGAAGAA 1920

GTAAAACCGA ATAATGGGAT GAAAATAAGA GTTCCTGATA TGATTAAAAT AATGTTGTTG 1980

GGAGTGATTG TTATGGGGAT GATGTAAAAT GAATGAAAAA AATGTTAATA GATTGAAATT 2040

GTGCATATAT CCATTGTTTA TATATAATAG AAATCTAAAT GAATGAATGA ATTAAAAAAT 2100

AAAGTTTTAA AAAAAAAAAA AAAAAA 2126

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 3521 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Pneumocystis carnii

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 146..3409
(D) OTHER INFORMATION: /product= "gp3"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

TAGACGATAT GAAGAGAGAA TTGAGTTAAA TCATTTGGGG AGACGCCCAG GAGTCGACTA 60

TTTTAGGAAA GGTGGGGATG TTTTTACTGA TGGTTATCCT CGTGGAGGTC ATTTGATCGA 120

GGATGAGTTG TCCGAAGAGG TGGCA ATG GCA CGG CCG GTT AAG AGG CAA GCA 172
Met Ala Arg Pro Val Lys Arg Gln Ala
1 5

GTA CAA GGA GCA CAA GAT GAG ATT GAT GAG AAA CAC CTT TTG GCT TTC 220
Val Gln Gly Ala Gln Asp Glu Ile Asp Glu Lys His Leu Leu Ala Phe
10 15 20 25

ATT GTG AAG GAC AAA TAT AAA GAA GAA CAA AAA TGC AAA GAA GAA CTC 268
Ile Val Lys Asp Lys Tyr Lys Glu Glu Gln Lys Cys Lys Glu Glu Leu
30 35 40

GAG AAA TAT TGT AAA GAG TTG AAG GAA GCA GAT AAA AAT CTA GAG AAT 316
Glu Lys Tyr Cys Lys Glu Leu Lys Glu Ala Asp Lys Asn Leu Glu Asn
45 50 55

GTG GAT GAT AAA GTT AAA GGG CTT TGT GAT GAT AAA AAA CGA GAC GAA 364
Val Asp Asp Lys Val Lys Gly Leu Cys Asp Asp Lys Lys Arg Asp Glu
60 65 70

```
AAA TGC AAA GAC GTG AAA AAA AAA GTT GAA GAT GAA TTA AAA GAT TTT                412
Lys Cys Lys Asp Val Lys Lys Lys Val Glu Asp Glu Leu Lys Asp Phe
     75                  80                  85

GAA GAG GAA CTT CAA AAA GTA TTG AAT AAT ATA AAA GAT GAA AAT TGC                460
Glu Glu Glu Leu Gln Lys Val Leu Asn Asn Ile Lys Asp Glu Asn Cys
 90                  95                 100                 105

GAA AAA TAT GAA GAA AAA TGT ATA CTT TTA GAA GAG ACG GAT TAT GAT                508
Glu Lys Tyr Glu Glu Lys Cys Ile Leu Leu Glu Glu Thr Asp Tyr Asp
                110                 115                 120

GTT ATT AAG GAT AAC TGT ATC GAG TTG AGG GAA GGA TGT TAC AAA TTG                556
Val Ile Lys Asp Asn Cys Ile Glu Leu Arg Glu Gly Cys Tyr Lys Leu
            125                 130                 135

AAG CGT GAA AAG GTG GCA GAG GAG CTT CTT CTG AGG GCG CTC GGA GGG                604
Lys Arg Glu Lys Val Ala Glu Glu Leu Leu Leu Arg Ala Leu Gly Gly
        140                 145                 150

GAT GCT AAA GAA GAA GCT AAA TGT AAA GGA AAG ATG AAT ACT GTT TGC                652
Asp Ala Lys Glu Glu Ala Lys Cys Lys Gly Lys Met Asn Thr Val Cys
    155                 160                 165

CCA GTG TTG AGC CGA GAA AGC GAC GAA TTG ATG TCT TTT TGC CTT GAT                700
Pro Val Leu Ser Arg Glu Ser Asp Glu Leu Met Ser Phe Cys Leu Asp
170                 175                 180                 185

TCT GCT AAA ACA TGT GGA GAT CTG AAA AAA AAA TTG GGT ACT GTT TGC                748
Ser Ala Lys Thr Cys Gly Asp Leu Lys Lys Lys Leu Gly Thr Val Cys
                190                 195                 200

GAG CCT TTA AAA AAA GAG CTT AAA GAT AAC GAA TTA GCG GAA AAG TGT                796
Glu Pro Leu Lys Lys Glu Leu Lys Asp Asn Glu Leu Ala Glu Lys Cys
            205                 210                 215

CAT GAA AGA CTT GAG AAA TGT CAT TTT TAC GGA GAA GCG TGT GAT GAT                844
His Glu Arg Leu Glu Lys Cys His Phe Tyr Gly Glu Ala Cys Asp Asp
        220                 225                 230

GCG AAA TGC AAG AAG TTT GAG GAG CAA TGC AAG GGA AAA AAT ATT ATA                892
Ala Lys Cys Lys Lys Phe Glu Glu Gln Cys Lys Gly Lys Asn Ile Ile
    235                 240                 245

TAT AAA GCG CCA GAA TCT GAT CTT AGT CCT GTC AAG CCG AGG GCG TCC                940
Tyr Lys Ala Pro Glu Ser Asp Leu Ser Pro Val Lys Pro Arg Ala Ser
250                 255                 260                 265

TTG TTG AGA AGT ATT GGG TTG GAT GAT GTG TAT AAA AAC GCG GAA AAA                988
Leu Leu Arg Ser Ile Gly Leu Asp Asp Val Tyr Lys Asn Ala Glu Lys
                270                 275                 280

CAT GGG ATT ATT ATT GGA AAA TCA GGA GTG GAT CTA CCA AGG AAG TCA               1036
His Gly Ile Ile Ile Gly Lys Ser Gly Val Asp Leu Pro Arg Lys Ser
            285                 290                 295

GGT ACA AAT TTC TGC AAG ATC TCT TTG CTA CTG TTG AGC AGA GAT GAG               1084
Gly Thr Asn Phe Cys Lys Ile Ser Leu Leu Leu Leu Ser Arg Asp Glu
        300                 305                 310

GAT AAG AAG GAA CCA GAT AAA AAG TGC ACT AAA GCG TTA GAA AAA TGT               1132
Asp Lys Lys Glu Pro Asp Lys Lys Cys Thr Lys Ala Leu Glu Lys Cys
    315                 320                 325

GAT GCC TCT AAG TAT TTG AAT ACT GAA TTG GAA AAG TTA TGT AAA GAT               1180
Asp Ala Ser Lys Tyr Leu Asn Thr Glu Leu Glu Lys Leu Cys Lys Asp
330                 335                 340                 345
```

```
GGA AAC AAA AAC GAA AAA TGC AAA AAA ATA TTA GAT GTA AAA GAA AGA      1228
Gly Asn Lys Asn Glu Lys Cys Lys Lys Ile Leu Asp Val Lys Glu Arg
                350                 355                 360

TGT ACA AAT CTC AAA TTA AAA CTT TAT CTG AAA GGA TTG TCT ACG GAA      1276
Cys Thr Asn Leu Lys Leu Lys Leu Tyr Leu Lys Gly Leu Ser Thr Glu
            365                 370                 375

TAT GAT GAT CAA GAA TCA GAT CCT TTA TCG TGG GGA CAG CTG CCA ACT      1324
Tyr Asp Asp Gln Glu Ser Asp Pro Leu Ser Trp Gly Gln Leu Pro Thr
        380                 385                 390

TTT TTT ATA AAA GGA GAG TGT GCA GAA CTT GAG TCG GAA TGT TTC TAT      1372
Phe Phe Ile Lys Gly Glu Cys Ala Glu Leu Glu Ser Glu Cys Phe Tyr
    395                 400                 405

TTA GAA AAG GCG TGT AAA GAT AAT AAT ATT GAT AAA GCG TGC CAA AAT      1420
Leu Glu Lys Ala Cys Lys Asp Asn Asn Ile Asp Lys Ala Cys Gln Asn
410                 415                 420                 425

GCA AGA GCA GCG TGC TAT AAA AAG GGA CAA GAC AGG ATG TTG AAT AAG      1468
Ala Arg Ala Ala Cys Tyr Lys Lys Gly Gln Asp Arg Met Leu Asn Lys
                430                 435                 440

TTC TTT CAA AAG GAA TTG AAG GGA AAG CTT GGT CAT GTA AGA TTT TAT      1516
Phe Phe Gln Lys Glu Leu Lys Gly Lys Leu Gly His Val Arg Phe Tyr
            445                 450                 455

AGC GAT CCT AAA GAT TGT AAA AAA TAT GTG GTA GAA AAC TGT ACA AAA      1564
Ser Asp Pro Lys Asp Cys Lys Lys Tyr Val Val Glu Asn Cys Thr Lys
        460                 465                 470

CTT GAT AAA AAA TAT CTT CCA CGA TGT CTT TAT CCT AAA GAA CTA TGT      1612
Leu Asp Lys Lys Tyr Leu Pro Arg Cys Leu Tyr Pro Lys Glu Leu Cys
    475                 480                 485

TAT GGG CTT TCA AAT GAT ATT TTT CTT CAA TCC AAA GAG TTA AGT GCG      1660
Tyr Gly Leu Ser Asn Asp Ile Phe Leu Gln Ser Lys Glu Leu Ser Ala
490                 495                 500                 505

CTT TTG GAT GAT CAA AGG GAT TTT CCA TTA AAA AAG GAT TGT GTT GAG      1708
Leu Leu Asp Asp Gln Arg Asp Phe Pro Leu Lys Lys Asp Cys Val Glu
                510                 515                 520

TTG AAG GAG AAG TGT GAT GAA CTT AGT AGT GAT TCA TTA TTG AAT TTA      1756
Leu Lys Glu Lys Cys Asp Glu Leu Ser Ser Asp Ser Leu Leu Asn Leu
            525                 530                 535

GAA AAG TGT ATA ACA TTG AAA AGA CGT TGT GAA TAC TTT AGA GTT TCA      1804
Glu Lys Cys Ile Thr Leu Lys Arg Arg Cys Glu Tyr Phe Arg Val Ser
        540                 545                 550

GAG GGA TTT AGA AAT GTA TTT TTA GAA AAA AAG GAT GAT TCG TTA ATG      1852
Glu Gly Phe Arg Asn Val Phe Leu Glu Lys Lys Asp Asp Ser Leu Met
    555                 560                 565

ACT CAG GAT AAC TGT ACA AAG GCA TTG CAT GAG AAA TGC CAT CAA TTA      1900
Thr Gln Asp Asn Cys Thr Lys Ala Leu His Glu Lys Cys His Gln Leu
570                 575                 580                 585

TAT AGG AGG AGA AAG AAT TCA TTT AGT GTT TCA TGT GCT TTA CCA GAA      1948
Tyr Arg Arg Arg Lys Asn Ser Phe Ser Val Ser Cys Ala Leu Pro Glu
                590                 595                 600

GAA ACA TGT AGT TAT ATG GTA TTC CAT ACA AGT CAA GAT TGT AGT AGT      1996
Glu Thr Cys Ser Tyr Met Val Phe His Thr Ser Gln Asp Cys Ser Ser
            605                 610                 615
```

```
TTA AAA GTC AAC ATC AAG AAT GAA AAA ATT CTA GAA AAA ATT GGA GAA      2044
Leu Lys Val Asn Ile Lys Asn Glu Lys Ile Leu Glu Lys Ile Gly Glu
        620                 625                 630

GAA ATT AAA AAA GCA AAT AAA AAT GAA GCC TTG GTT GAA GAA CTC TGC      2092
Glu Ile Lys Lys Ala Asn Lys Asn Glu Ala Leu Val Glu Glu Leu Cys
    635                 640                 645

ACA ACA TGG GGC CGA CAT TGT CAC CAA CTT ATG GAG AAT TGT CCG GAT      2140
Thr Thr Trp Gly Arg His Cys His Gln Leu Met Glu Asn Cys Pro Asp
650                 655                 660                 665

GAC TTG AAA AAA AAA GAG AAT GGC AAT GGC AAT GAT CAT AAC TGC GAA      2188
Asp Leu Lys Lys Lys Glu Asn Gly Asn Gly Asn Asp His Asn Cys Glu
                670                 675                 680

GCA CTC CAA GAA AAA TGC AAT AAA ACC TTT GAA AAG TTG AAA TTA GAG      2236
Ala Leu Gln Glu Lys Cys Asn Lys Thr Phe Glu Lys Leu Lys Leu Glu
            685                 690                 695

GAG GAG CTG AGT CAT CTG TTG AAA GGC AGT TTA AAG GAT GAT AAA TGT      2284
Glu Glu Leu Ser His Leu Leu Lys Gly Ser Leu Lys Asp Asp Lys Cys
        700                 705                 710

AAA GAA GCA TTA GGA AAG CGT TGC ACT GAG TTG GAA AAG AAT GAA GCA      2332
Lys Glu Ala Leu Gly Lys Arg Cys Thr Glu Leu Glu Lys Asn Glu Ala
    715                 720                 725

TTC AAA ACT CTG TAT GGT AAA TGT GAT GAT AAT ACC AAG GAA AAT GTT      2380
Phe Lys Thr Leu Tyr Gly Lys Cys Asp Asp Asn Thr Lys Glu Asn Val
730                 735                 740                 745

TGC AAA AAA TTA GTT GAT AAA GTA AAA AAG AGA TGC CCT ACT TTA AAA      2428
Cys Lys Lys Leu Val Asp Lys Val Lys Lys Arg Cys Pro Thr Leu Lys
                750                 755                 760

GAC GAA CTG GAG AAT GCG AAA AAA GAG TTG ACA AAG ATG AAG AAT GAG      2476
Asp Glu Leu Glu Asn Ala Lys Lys Glu Leu Thr Lys Met Lys Asn Glu
            765                 770                 775

TAC GAT GAT CTC AAA AAG GCG GCA GAA AAA TCT ACG GAG GCA GCT AAG      2524
Tyr Asp Asp Leu Lys Lys Ala Ala Glu Lys Ser Thr Glu Ala Ala Lys
        780                 785                 790

TTA TTG CTA TCA AGA CCT AGA CAA ACT GTA ATG CCA AAT GCG CAG AAT      2572
Leu Leu Leu Ser Arg Pro Arg Gln Thr Val Met Pro Asn Ala Gln Asn
    795                 800                 805

GGC AGT GAT TCT ACA CTA GTA CCA CCA CCA CCA CAA GCA CCA GCA GGG      2620
Gly Ser Asp Ser Thr Leu Val Pro Pro Pro Pro Gln Ala Pro Ala Gly
810                 815                 820                 825

CCA CCA CCA CCA GGG TCA CCA CCA CCA CCA CCA TCA CAA AAT GGA ACG      2668
Pro Pro Pro Pro Gly Ser Pro Pro Pro Pro Pro Ser Gln Asn Gly Thr
                830                 835                 840

CCA GGC ACA CCA GGT GGA GAA ACA GGC GCA TCA GGT GGA ACA CCA GGC      2716
Pro Gly Thr Pro Gly Gly Glu Thr Gly Ala Ser Gly Gly Thr Pro Gly
            845                 850                 855

ACA CCA GGC ACA CCA GGC ACA CCA GGC ACA CCA GGT GGA ATG ATG AAG      2764
Thr Pro Gly Thr Pro Gly Thr Pro Gly Thr Pro Gly Gly Met Met Lys
        860                 865                 870

TAT GCA AAA CTT GGA CTC GTT AAA AGA ACG TAT GTA GAT GGA GGT GTA      2812
Tyr Ala Lys Leu Gly Leu Val Lys Arg Thr Tyr Val Asp Gly Gly Val
    875                 880                 885
```

```
TCA GAA GTA GAG GTC AAA GCA TTT GAT GCA ACG ACG ATA GCA TTG GAA      2860
Ser Glu Val Glu Val Lys Ala Phe Asp Ala Thr Thr Ile Ala Leu Glu
890                 895                 900                 905

TTG TAT TTG GAA TTG AAA GAA GAA TGT AAA GCT TTA GAA TTA GAT TGC      2908
Leu Tyr Leu Glu Leu Lys Glu Glu Cys Lys Ala Leu Glu Leu Asp Cys
                910                 915                 920

GGT TTT AAA GAG GAT TGT CCA GAT ACT AAA CAA GCT TGC GAA AAT ATA      2956
Gly Phe Lys Glu Asp Cys Pro Asp Thr Lys Gln Ala Cys Glu Asn Ile
            925                 930                 935

GAC ACT TTA TGT AAA CTG GAA CCA TTA GAA ATT AAG CCT CAT CAT ACA      3004
Asp Thr Leu Cys Lys Leu Glu Pro Leu Glu Ile Lys Pro His His Thr
        940                 945                 950

GAG AAA ATA ACA GAA ACA AAG ACG GAA ACG AAG ACG GAA ACA AAG ACG      3052
Glu Lys Ile Thr Glu Thr Lys Thr Glu Thr Lys Thr Glu Thr Lys Thr
    955                 960                 965

GAA ACA AAG ACT GAT GGC AAG GCT GAT GAA AAG ACC GTT GAG AAG ACT      3100
Glu Thr Lys Thr Asp Gly Lys Ala Asp Glu Lys Thr Val Glu Lys Thr
970                 975                 980                 985

GTT ACA GAA ACC AAG TCA GTA GGT GGA GGA AAA GTA ACA GAA GAG TGT      3148
Val Thr Glu Thr Lys Ser Val Gly Gly Gly Lys Val Thr Glu Glu Cys
                990                 995                 1000

ACA ATG ATA CAA ACA ACA GAT ACA TGG GTG ACG AGT ACG TCA TTG CAT      3196
Thr Met Ile Gln Thr Thr Asp Thr Trp Val Thr Ser Thr Ser Leu His
            1005                1010                1015

ACG AGT ACG ACA ACG AGT ACG TCA ACG GTG ACG TCG ACA GTG ACG TTG      3244
Thr Ser Thr Thr Thr Ser Thr Ser Thr Val Thr Ser Thr Val Thr Leu
        1020                1025                1030

ACT TCG ATG CGC AAG TGC AAG CCT ACC AAA TGT ACC ACC GAT TCA AGC      3292
Thr Ser Met Arg Lys Cys Lys Pro Thr Lys Cys Thr Thr Asp Ser Ser
    1035                1040                1045

AAA GAG ACA CAG AAA GAA GAA GAT GAT GAA GAA GTG AAA CCG AAT GAG      3340
Lys Glu Thr Gln Lys Glu Glu Asp Asp Glu Glu Val Lys Pro Asn Glu
1050                1055                1060                1065

GGA ATG AAA ATA AGA GTT CCT GAT ATG ATT AAA ATA ATG TTG TTG GGA      3388
Gly Met Lys Ile Arg Val Pro Asp Met Ile Lys Ile Met Leu Leu Gly
                1070                1075                1080

GTG ATT GTT ATG GGG ATG ATG TAAATGAATG AAAAAAATGT TAATAGATTG         3439
Val Ile Val Met Gly Met Met
            1085

AAAATGTGCA TATATCCATT GTTTATATAT AATAAAAATG TTAAAGAATG AAATGAAAAA    3499

AAAAAAAAAA AAAAAAAAAA AA                                             3521
```

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1088 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

```
Met Ala Arg Pro Val Lys Arg Gln Ala Val Gln Gly Ala Gln Asp Glu
 1               5                  10                  15

Ile Asp Glu Lys His Leu Leu Ala Phe Ile Val Lys Asp Lys Tyr Lys
            20                  25                  30

Glu Glu Gln Lys Cys Lys Glu Glu Leu Glu Lys Tyr Cys Lys Glu Leu
        35                  40                  45

Lys Glu Ala Asp Lys Asn Leu Glu Asn Val Asp Asp Lys Val Lys Gly
    50                  55                  60

Leu Cys Asp Asp Lys Lys Arg Asp Glu Lys Cys Lys Asp Val Lys Lys
65                  70                  75                  80

Lys Val Glu Asp Glu Leu Lys Asp Phe Glu Glu Glu Leu Gln Lys Val
                85                  90                  95

Leu Asn Asn Ile Lys Asp Glu Asn Cys Glu Lys Tyr Glu Glu Lys Cys
            100                 105                 110

Ile Leu Leu Glu Glu Thr Asp Tyr Asp Val Ile Lys Asp Asn Cys Ile
        115                 120                 125

Glu Leu Arg Glu Gly Cys Tyr Lys Leu Lys Arg Glu Lys Val Ala Glu
    130                 135                 140

Glu Leu Leu Leu Arg Ala Leu Gly Gly Asp Ala Lys Glu Glu Ala Lys
145                 150                 155                 160

Cys Lys Gly Lys Met Asn Thr Val Cys Pro Val Leu Ser Arg Glu Ser
                165                 170                 175

Asp Glu Leu Met Ser Phe Cys Leu Asp Ser Ala Lys Thr Cys Gly Asp
            180                 185                 190

Leu Lys Lys Lys Leu Gly Thr Val Cys Glu Pro Leu Lys Lys Glu Leu
        195                 200                 205

Lys Asp Asn Glu Leu Ala Glu Lys Cys His Glu Arg Leu Glu Lys Cys
    210                 215                 220

His Phe Tyr Gly Glu Ala Cys Asp Asp Ala Lys Cys Lys Lys Phe Glu
225                 230                 235                 240

Glu Gln Cys Lys Gly Lys Asn Ile Ile Tyr Lys Ala Pro Glu Ser Asp
                245                 250                 255

Leu Ser Pro Val Lys Pro Arg Ala Ser Leu Leu Arg Ser Ile Gly Leu
            260                 265                 270

Asp Asp Val Tyr Lys Asn Ala Glu Lys His Gly Ile Ile Ile Gly Lys
        275                 280                 285

Ser Gly Val Asp Leu Pro Arg Lys Ser Gly Thr Asn Phe Cys Lys Ile
    290                 295                 300

Ser Leu Leu Leu Leu Ser Arg Asp Glu Asp Lys Lys Glu Pro Asp Lys
305                 310                 315                 320

Lys Cys Thr Lys Ala Leu Glu Lys Cys Asp Ala Ser Lys Tyr Leu Asn
                325                 330                 335

Thr Glu Leu Glu Lys Leu Cys Lys Asp Gly Asn Lys Asn Glu Lys Cys
            340                 345                 350
```

```
Lys Lys Ile Leu Asp Val Lys Glu Arg Cys Thr Asn Leu Lys Leu Lys
        355                 360                 365

Leu Tyr Leu Lys Gly Leu Ser Thr Glu Tyr Asp Asp Gln Glu Ser Asp
    370                 375                 380

Pro Leu Ser Trp Gly Gln Leu Pro Thr Phe Phe Ile Lys Gly Glu Cys
385                 390                 395                 400

Ala Glu Leu Glu Ser Glu Cys Phe Tyr Leu Glu Lys Ala Cys Lys Asp
                405                 410                 415

Asn Asn Ile Asp Lys Ala Cys Gln Asn Ala Arg Ala Ala Cys Tyr Lys
            420                 425                 430

Lys Gly Gln Asp Arg Met Leu Asn Lys Phe Phe Gln Lys Glu Leu Lys
        435                 440                 445

Gly Lys Leu Gly His Val Arg Phe Tyr Ser Asp Pro Lys Asp Cys Lys
    450                 455                 460

Lys Tyr Val Val Glu Asn Cys Thr Lys Leu Asp Lys Lys Tyr Leu Pro
465                 470                 475                 480

Arg Cys Leu Tyr Pro Lys Glu Leu Cys Tyr Gly Leu Ser Asn Asp Ile
                485                 490                 495

Phe Leu Gln Ser Lys Glu Leu Ser Ala Leu Leu Asp Asp Gln Arg Asp
            500                 505                 510

Phe Pro Leu Lys Lys Asp Cys Val Glu Leu Lys Glu Lys Cys Asp Glu
        515                 520                 525

Leu Ser Ser Asp Ser Leu Leu Asn Leu Glu Lys Cys Ile Thr Leu Lys
    530                 535                 540

Arg Arg Cys Glu Tyr Phe Arg Val Ser Glu Gly Phe Arg Asn Val Phe
545                 550                 555                 560

Leu Glu Lys Lys Asp Asp Ser Leu Met Thr Gln Asp Asn Cys Thr Lys
                565                 570                 575

Ala Leu His Glu Lys Cys His Gln Leu Tyr Arg Arg Arg Lys Asn Ser
            580                 585                 590

Phe Ser Val Ser Cys Ala Leu Pro Glu Glu Thr Cys Ser Tyr Met Val
        595                 600                 605

Phe His Thr Ser Gln Asp Cys Ser Ser Leu Lys Val Asn Ile Lys Asn
    610                 615                 620

Glu Lys Ile Leu Glu Lys Ile Gly Glu Glu Ile Lys Lys Ala Asn Lys
625                 630                 635                 640

Asn Glu Ala Leu Val Glu Glu Leu Cys Thr Thr Trp Gly Arg His Cys
                645                 650                 655

His Gln Leu Met Glu Asn Cys Pro Asp Asp Leu Lys Lys Lys Glu Asn
            660                 665                 670

Gly Asn Gly Asn Asp His Asn Cys Glu Ala Leu Gln Glu Lys Cys Asn
        675                 680                 685

Lys Thr Phe Glu Lys Leu Lys Leu Glu Glu Glu Leu Ser His Leu Leu
    690                 695                 700
```

Lys Gly Ser Leu Lys Asp Asp Lys Cys Lys Glu Ala Leu Gly Lys Arg
705 710 715 720

Cys Thr Glu Leu Glu Lys Asn Glu Ala Phe Lys Thr Leu Tyr Gly Lys
725 730 735

Cys Asp Asp Asn Thr Lys Glu Asn Val Cys Lys Lys Leu Val Asp Lys
740 745 750

Val Lys Lys Arg Cys Pro Thr Leu Lys Asp Glu Leu Glu Asn Ala Lys
755 760 765

Lys Glu Leu Thr Lys Met Lys Asn Glu Tyr Asp Asp Leu Lys Lys Ala
770 775 780

Ala Glu Lys Ser Thr Glu Ala Ala Lys Leu Leu Leu Ser Arg Pro Arg
785 790 795 800

Gln Thr Val Met Pro Asn Ala Gln Asn Gly Ser Asp Ser Thr Leu Val
805 810 815

Pro Pro Pro Pro Gln Ala Pro Ala Gly Pro Pro Pro Pro Gly Ser Pro
820 825 830

Pro Pro Pro Pro Ser Gln Asn Gly Thr Pro Gly Thr Pro Gly Gly Glu
835 840 845

Thr Gly Ala Ser Gly Gly Thr Pro Gly Thr Pro Gly Thr Pro Gly Thr
850 855 860

Pro Gly Thr Pro Gly Gly Met Met Lys Tyr Ala Lys Leu Gly Leu Val
865 870 875 880

Lys Arg Thr Tyr Val Asp Gly Gly Val Ser Glu Val Glu Val Lys Ala
885 890 895

Phe Asp Ala Thr Thr Ile Ala Leu Glu Leu Tyr Leu Glu Leu Lys Glu
900 905 910

Glu Cys Lys Ala Leu Glu Leu Asp Cys Gly Phe Lys Glu Asp Cys Pro
915 920 925

Asp Thr Lys Gln Ala Cys Glu Asn Ile Asp Thr Leu Cys Lys Leu Glu
930 935 940

Pro Leu Glu Ile Lys Pro His His Thr Glu Lys Ile Thr Glu Thr Lys
945 950 955 960

Thr Glu Thr Lys Thr Glu Thr Lys Thr Glu Thr Lys Thr Asp Gly Lys
965 970 975

Ala Asp Glu Lys Thr Val Glu Lys Thr Val Thr Glu Thr Lys Ser Val
980 985 990

Gly Gly Gly Lys Val Thr Glu Glu Cys Thr Met Ile Gln Thr Thr Asp
995 1000 1005

Thr Trp Val Thr Ser Thr Ser Leu His Thr Ser Thr Thr Thr Ser Thr
1010 1015 1020

Ser Thr Val Thr Ser Thr Val Thr Leu Thr Ser Met Arg Lys Cys Lys
1025 1030 1035 1040

Pro Thr Lys Cys Thr Thr Asp Ser Ser Lys Glu Thr Gln Lys Glu Glu
1045 1050 1055

Asp Asp Glu Glu Val Lys Pro Asn Glu Gly Met Lys Ile Arg Val Pro
1060 1065 1070

Asp Met Ile Lys Ile Met Leu Leu Gly Val Ile Val Met Gly Met Met
1075 1080 1085

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(iii) HYPOTHETICAL: NO

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..20
(D) OTHER INFORMATION: /label= oligonucleotide
/note= "primer JK58"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

TTAACCGGCC GTGCCATTGC 20

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 36 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..35
(D) OTHER INFORMATION: /label= oligonucleotide
/note= "primer ANC"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GACTGCATGC GGAAGCTTGG ATCCCCCCCC CCCCCC 36

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..24

(D) OTHER INFORMATION: /label= oligonucleotide
/note= "primer AN"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GACTGCATGC GGAAGCTTGG ATCC 24

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..20
(D) OTHER INFORMATION: /label= oligonucleotide
/note= "5' primer corresponding to first 20 bases
of GP3 mRNA"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

TTTTTCTAAT AGACGATATG 20

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..20
(D) OTHER INFORMATION: /label= oligonucleotide
/note= "3' primer corresponding to positions 77-96
of GP3"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

GATCTCCACA TGTTTTAGCA 20

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: -

(B) LOCATION: 1..30
(D) OTHER INFORMATION: /label= oligonucleotide
/note= "hybridization probe MSG1"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GCAGAACTTG AGTCGGAATG TTTYTATTTA 30

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..30
(D) OTHER INFORMATION: /label= oligonucleotide
/note= "hybridization probe MSG2"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

AAAATATCTT CCACGATGTC TTTATCCTAA 30

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..30
(D) OTHER INFORMATION: /label= oligonucleotide
/note= "hybridization probe MSG3"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GAAAATAAAG ATAAGAGATA CCTTCCAAAG 30

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..30

(D) OTHER INFORMATION: /label= oligonucleotide
/note= "hybridization probe DHPS1"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

TTGATCACGA TATTAAGCCA GTTTTGCCAT

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 15 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(v) FRAGMENT TYPE: internal

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

Glu Leu Lys Gly Lys Leu Gly His Val Arg Phe Tyr Ser Asp Pro
1 5 10 15

What is claimed is:

1. A DNA molecule encoding a mammalian Pneumocystis carinii major surface glycoprotein or an allelic variation thereof.

2. The DNA molecule according to claim 1 where the mammal is a rat.

3. A DNA molecule encoding the gene for the major surface glycoprotein of P. carinii as shown in SEQ ID NO: 8.

4. A DNA molecule encoding a portion of the gene for the major surface glycoprotein of P. carinii in a cDNA selected from the group consisting of SEQ ID NO: 1, SEQ ID NO: 2, SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6 and SEQ ID NO: 7.

5. A DNA molecule according to claim 1 where the mammal is a human.

6. A DNA molecule encoding a mammalian Pneumocystis carinii major surface glycoprotein which is a composite of a multiple gene family or is a synthetic construction representing a consensus sequence analysis of a multiple gene family.

7. A method of obtaining a DNA molecule encoding a mammalian P. carinii major surface glycoprotein which comprises screening a cDNA library of P. carinii with an antibody to said major surface glycoprotein to identify positive clones encoding for gp116 and using at least one of said clones or an oligonucleotide probe based on said clones to reveal the presence of multiple genes encoding for said major surface glycoprotein.

8. A mammalian Pneumocystis carinii major surface glycoprotein having the amino acid sequence as shown in SEQ ID NO: 9.

9. A mammalian Pneumocystis carinii major surface glycoprotein produced from the expression of a DNA sequence which is a composite of a multiple gene family encoding for said major surface glycoprotein or a synthetic construction representing a consensus sequence analysis of a multiple gene family.

10. The major surface glycoprotein according to claim 8 where the mammal is human.

11. A mammalian Pneumocystis carinii major surface protein having a molecular weight of about 122977 or allelic variations thereof.

12. A vaccine comprising a therapeutically effective amount of a mammaian Pneumocystis carinii major surface glycoprotein or a polypeptide derived therefrom capable of eliciting an immune response to said glycoprotein, and a pharmaceutically acceptable parenteral vehicle.

```
PC3   MARPVKRQAKVVQGAQGAQDDIKEEHLLAFIAKKEYSNEDKCKQELKKYCEELKEADGKF-NVNDKVKELCGGGDEAKRDKKCKDLKDKVEDELENFDDE   99
PC5   ------------------------------------------------------------------------------------------EEELTTFEGD   10
GP3   MARPVKRQA-----VQGAQDEIDEKHLLAFIVKDKYKEEQKCKEELEKYCKELKEADKNLENVDDKVKGLC---DDKKRDEKCKDVKKKVEDELKDFEEE   92
GP22  ----------------------------------------------------------------------------------------------------    0
GP46  ----------------------------------------------------------------------------------------------------    0
GP14  ----------------------------------------------------------------------------------------------------    0
PC14  ----------------------------------------------------------------------------------------------------    0
      *********      ***** * * ****** *   *  * *** ** *** ******      ** **** **    *  *** **** * *** **  *

                           Peptide 1: LREGCYELK
PC3   LQEALKD-IKDENCEKHEEKCILLEDTGYSEDIKKNCVKLREGCYKLKRKKVAEELLLRALGGDAKDEAKCKEKMKTVCPMLSRESDELMFFCLDSDGTC  198
PC5   LDTALKNGIKDEDCEKHEEKCILLEEAD-PNSLKEKCVKLREGCYELKREKVAEELLFRALGGDAKEDGKCKGKMNTVCPVLSRESDELMTFCLDPDGTC  109
GP3   LQKVLNN-IKDENCEKYEEKCILLEETDY-DVIKDNCIELREGCYKLKREKVAEELLLRALGGDAKEEAKCKGKMNTVCPVLSRESDELMSFCLDSAKTC  190
GP22  ----------------------------------------------------------------------------------------------------    0
GP46  ----------------------------------------------------------------------------------------------------    0
GP14  ----------------------------------------------------------------------------------------------------    0
PC14  ----------------------------------------------------------------------------------------------------    0
      *   *   **** *** ********         *  * ******* *** ******* ********    *** ** **** ********* ****   **

PC3   KALKTKSEEVCLPLKEKLKDGELKEKCHERLEKCHFYKEACTETKCDEDMKQCKEKGFTYKAPESDFSPVKPKASLLRSIGLDDVYKKAEKEGIIIGKSG  298
PC5   GELKTKLGEVCKPLETELNEKS-SEKCHERLEKCHFYKEACGNTKCKEDKTKCEEKGFTYKAPESDFSPVKPKASLLRSIGLEDVYKNAEKHGIIIGKSG  208
GP3   GDLKKKLGTVCEPLKKELKDNELAEKCHERLEKCHFYGEACDDAKCKKFEEQCKGKNIIYKAPESDLSPVKPRASLLRSIGLDDVYKNAEKHGIIIGKSG  290
GP22  ----------------------------------------------------------------------------------------------------    0
GP46  ----------------------------------------------------------------------------------------------------    0
GP14  ----------------------------------------------------------------------------------------------------    0
PC14  ----------------------------------------------------------------------------------------------------    0
        ** *   ** **    *        ************* ***    **        *  *    ******* ***** ********* **** *** ********

PC3   KQCKEKGFTYKAPESDFSPVKPKAALLRSIGLDDVYKKAEKEGIIIGKSGVDLPRKSGTKFLQDLLLLLSRDEN--DAGKKCGKALGKCETSKYLNTDLM  346
PC5   TKCEEKGFTYKAPESDFSPVKPKASLLRSIGLEDVYKNAEKHGIIIGKSGVDLPRKSGTKFLQDLLLVLSRDEN--DAGKKCGKALGKCDASKYLDHNLK  256
GP3   EQCKGKNIIYKAPESDLSPVKPRASLLRSIGLDDVYKNAEKHGIIIGKSGVDLPRKSGTNFCKISLLLLSRDEDKKEPDKKCTKALEKCDASKYLNTELE  340
GP22  ----------------------------------------------------------------------------------------------------    0
GP46  ----------------------------------------------------------------------------------------------------    0
GP14  ----------------------------------------------------------------------------------------------------    0
PC14  ----------------------------------------------------------------------------------------------------    0
        *  *   ******* ***** * ******* **** *** ***************** *     ** *****     *** *** **  ****    *

PC3   ELCKDADKENKCKKKLDV--KERCTKLKLNLYVKGLSTEFKEDKKSHLLSWGQLPTLFTKGECAELESECFYLENACKDNEIGEACQNLRSACYKKGQDR  444
PC5   ELCNDGKKNDKCKELLDVNVKERCTKLKLNLYVKGLSTKFEKAEKSDLLSWGQLPTLFTKGECAELESECFYLENACKDNKIDEACQNARAACYKKGQDR  356
GP3   KLCKDGNKNEKCKKILD--VKERCTNLKLKLYLKGLSTEYDDQE-SDPLSWGQLPTFFIKGECAELESECFYLEKACKDNNIDKACQNARAACYKKGQDR  437
GP22  -------------------------------------------------------------------AELVSECFYLEKACKDNKIDQACQNVRAACYKMGQNR   37
GP46  ----------------------------------------------------------------------------------------------------    0
GP14  ---------------------------------------------------------------------AELESECFYLKKACKDKEIDEACQNARAACYKVGKDR   37
PC14  ----------------------------------------------FPKLLSWGQLPTLFIKGECAELESECFYLENACTNK-IGEACQNVRSACYKKGQDR   55
      ** *  *  ***  **     ***** *** ** *****       ******** * ******* ******   **    *  **** * **** *  *

Peptide 2: ELRGNLGLVRFYSD                                                Peptide 3: ELSSLLDDQRDFPLEKDC
PC3   MLNKFFQKELKGKLGHVRFYSD-PKDCKKYVVENCTKL-KKDKRYLSKCLYPKELCYGLSNDIFLQSKELSSLLDDQRDFPFEKDCLELGEKCDQLSSDS  542
PC5   MLNKFFQKELRGNLGLVRFYSD-PEECKKSVVGNCTKL-KEDSRYLSKCLYPKELCYALSNDIFLQSKELSSLLDDQRDFPLEKDCLELVEKCDELSSDS  454
GP3   MLNKFFQKELKGKLGHVRFYSD-PKDCKKYVVENCTKL---DKKYLPRCLYPKELCYGLSNDIFLQSKELSALLDDQRDFPLKKDCVELKEKCDELSSDS  533
GP22  MLNMLFREGLKENSERIKYYDENPQKCQEFVVGSCTKL----KKYLPQCLYPKELCYAVSDDIFLQSKELGVLLDDQRDFPLEEDCLELKEKCAQLETYS  133
GP46  MLSTLFRKESKGESGHKRYYN-HPEECQKSVVKDCKKLENKDKRYLPKCLYPKELCYMLSDDIFLQSKELGALLDDQRDFPLEKHCVELKEKCDELETYS   99
GP14  MLSTLFRKESKGESGHKRYYN-HPEECQKSVVKDCKKLENKDKRYLPKCLYPKELCYMLSDDIFLQSKELGALLDDQRDFPLEKHCVELKEKCDELETYS  136
PC14  MLNTLFREEMKGKLANIKYFNDT-ESCKKSVAKKCAEL---DKRYLSKCLYPKRLCYVLSDDIFLQSKELSVLLDDQRDFPLEKDCVELGEKCDELGSDS  151
      **    *               *     *    *    *       **  ***** ***  *  *********  *********         * ** ***   *    *
```

FIGURE 1A

```
PC3   LLNLEKCITLKRRCEYFDVTERFRKVFLEKKDDSLMIQENCTKALHEKCNTLYKRRKN------------------------------------------   600
PC5   LLNLEKCITLKRRCEYFKVTEGFRKVFLKKK----------------------------------------------------------------------   489
GP3   LLNLEKCITLKRRCEYFRVSEGFRNVFLEKKDDSLMTQDNCTKALHEKCHQLYRRRKNSFSVSCALPEETCSYMVFHTSQDCSSLKVNIKNEKILEKIGE   633
GP22  LLNLEKCITLKRRCEYFKVTEGFRNVFLKREDDSLK-KENCTKALQEKCDALSRKRRNPFGFSCALREETCEYMVARTKDECFYLKDNMENEEILKEIEE   232
GP46  NSNSQKCATLRRRCKYLRVSEGFRNVFLKREDDSLK-KENCTKALQEKCDALSRKRRNPFGFSCALREETCEYMVYRTKDECFYLSGNMEDEKIVEEIGK   199
GP14  HSNSEKCITLRRRCEYLRVSEEFRKVFLKRKDHALYNEQNCTEVLQEKCNTLYRRRKNSFSVSCALPGETCEYMVYRTKDECFYLSGNMEDEKIVEEIGK   236
PC14  LLNLEKCITLKRRCEYFKVTERFRKVFLERKDHSLYDEQNCTKALHEKCEALFRKRRNPFEFSCALQEETCQRMVYHTTQDCIYLKDNIKNKKILEQIG-   250
        *  ** ** *** *  * * ** ***    *  *   ***  * ***  *    * * *  ****  ***  **  *    *  *  *     *  *

PC3   ----------------------------------------------------------------------------------------------------   600
PC5   ----------------------------------------------------------------------------------------------------   489
GP3   EIKK--ANKNEALVEELCTTWGRHCHQLMENCPDDLKKKEN-GNGNDHNCEALQEKCNKTFEKLKLEEELSHLLKGSLK-DDKCKEALGKRCTELEKNEA   729
GP22  KAKKDNANRNETLVEELCTTWGRHCHQLVGNCPEQLKKKKKKDDNKDHNCDKLEEKCSDTFKRLKLEEELTHLLKGSLKSEDECKKTLGEHCPELQKNDT   332
GP46  K------KANETALEELCTTWGRHCHQLMENCPDDLKKKEN-GNDNDHNCEELDEKCSDTFKRLKLEEELTHLLKGSLKDKDECKKTLEKRCTELQNNET   292
GP14  K------KANETALEELCTTWGRHCHQLMENCPDKLKK--E-SDNRDHNCDKLEEKCSDTFKKLKLKEELTHLLKGSLNDKKKCTETLGKNCTELQKNDT   327
PC14  --KVKQDKSKEAEVEELCTTWGKYCHQLMENCPDKLKKKKKKDNDNNQNCEELEKKCTDTFKKLELKDELTHLLKGSLKDKEKCKVTLGQRCPELKNNDT   348
                *    ********  ****  ***  ***            **  *  **  **  *  *  ** *******     *   *   * **  *

PC3   ----------------------------------------------------------------------------------------------------   600
PC5   ----------------------------------------------------------------------------------------------------   489
GP3   FKTLYGKCDDNTKENVCKKLVDKVKKRCPTLKDELENAKKELTKMKNEYDDLKKAAEKSTEAAKLLLSRPRQTVMPNAQNGSDSTLVPPPPQAPAGPPPP   829
GP22  FKTLYGKCEENEKGTVCKKLVKKVQERCPTLKTDLEKAKKELKDKKDEYDNVKQAAKESTEKAKLLLSKPRQTVTPNAQNGSASGPVPAPAAPPAAPEAP   432
GP46  FKNLLSYCGENDKGTVCEKLVEKLKKRCPTLKDGLNKAKDELTKMKKEYDALKKAAEESTKEASLLLSRPRQTVMPSAQNGSASEQVLQPVQPESGSSSG   392
GP14  FKILLSDC-KDSLENVCTKLVEKVQKRCPALKTDLEEAKKELKVKKEEYDALKKAAEESRNKASLLLSRSKQAVTPSGQNGS------DSVPAQVQPAPA   420
PC14  FKILLTNCEDS-LENVCAELVKKVQKKCPTLKDELNKAKDELTKMKTEYENAKKAAEESTNKASLLLSKSGKAAMPTAQNGSASAPPSAPAESGSSPASG   447
      ** *    *         **  ** *    ** **  *  ** **     * **     * **    *  * ****        *  ****

PC3   ----------------------------------------------------------------------------------------------------   600
PC5   ----------------------------------------------------------------------------------------------------   489
GP3   GSPPPPPSQNGTPGTPGGETGASGGTPGTPGTPGTPGTPGGMMKYAKLGLVKRTYVDGGVSEVEVKAFDATTIALELYLELKEECKALELDCGFKEDCPD   929
GP22  AQPPPPAGQPSG-------ETSNVPGKTPSKEAGTPNTTDETTKNPSLGLVKRAYVEGGVSEAEVKAFDATTIALELYLELKEECSALQLDCGFRKDCSS   525
GP46  SPSSPPGPPSAPPQNGTPATPGGAPGTPSSGTTGP----------AKLGLVKRAYVAEGVSEAEVKAFDATTIALELYLELKEECKALELDCGFKEDCKE   482
GP14  GPPSAPGSPSSPPSQNGTPGAPDGTTDTAGGTTNN----------AKLGLVKRAYVDEGVSEAEVKAFNATTIALELYLELKEECSALQLDCGFKEDCPD   510
PC14  SPPAS---------EPSTNGKVDTPAGGSGTQDKTSDASGQTTKYTKLGLVKRAYVAEGVSEEEVKAFDATTVALELYLELKEECNALELDCGFKEDCEE   538
                                                      ****** **  ****  ***** *** ************ ** *****  **

PC3   ----------------------------------------------------------------------------------------------------   600
PC5   ----------------------------------------------------------------------------------------------------   489
GP3   TKQACENIDTLCK-LEPLEIKPHHTEKITETKTETKTETKTETKTDGKADEKTVEKTVTETKSVGGGKVTEECTMIQTTDTWVTSTSLHTSTTTSTSTVT   1028
GP22  VEGVCKEIDKLCE-LEPLKVTPHHTETITNKVTETKTETKTETKTDDKADEKTGTKTVTETKTIGGGKVTEECTMVQTTDTWITRTSLHTSTTTSTSTVT   624
GP46  TEPACKEIEKLCK-LEALKVAPHHTETITNKVTETQTETKTVEKVDDKADVKTVEKTVTVTKPGSGEKVTEECTMIQTTDTWVTSTSLHTSTTTSTSTVT   581
GP14  TKQACKEIEELCK-LEALKVAPHHTETITETKTETKTETKMETKTDDKADEKTGTKTVTETKTIGGGKVTEECTLVKTTDTWVTSTSLHTSTTTSTSTVT   609
PC14  SKPACKEIEELCKGIESLKVTPHHTETQKEISTTTTTTTTTTTTTTT----TTTTTTTTKPGSGGKVTEECTMIQTTDTWVTRTSLHTSTTTSTSTVT   634
        *   *  **   *  *   *****      * * * *        *    * * *     * *******   ***** * ****************

PC3   ------------------------------------------------------------   600
PC5   ------------------------------------------------------------   489
GP3   STVTLTSMRKCKPTKCTTDSSKETQKE-EDDEEVKPNEGMKIRVPDMIKIMLLGVIVMGMM   1088
GP22  STVTLTSMRKCKPTKCTTDSTKETQKE-EDDEEVKPNEGMKIRVPDMIKIMLLGVIVTGMM   684
GP46  STVTLTSMRKCKPTKCTTDSSRETDKGGEGEEDVKPNEGMKIRVPDMIKIMLLGVIVMGMM   642
GP14  STVTLTSMRKCKPTKCTTDSTKETQK--EEDEEVKPNNGMKIRVPDMIKIMLLGVIVMGMM   668
PC14  STVTLTSMRKCKPTKCTTDSSKETEEGGKEEEEVKPNDGMKIRVPDMIKIILLGVIVMGMM   695
      ******************** **        * **** *********** ****** ***
```

FIGURE 1B

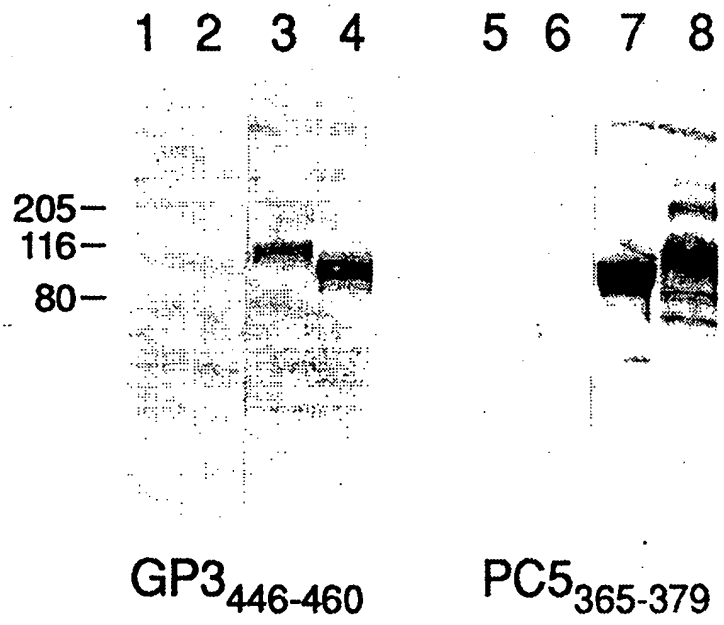

FIG. 2

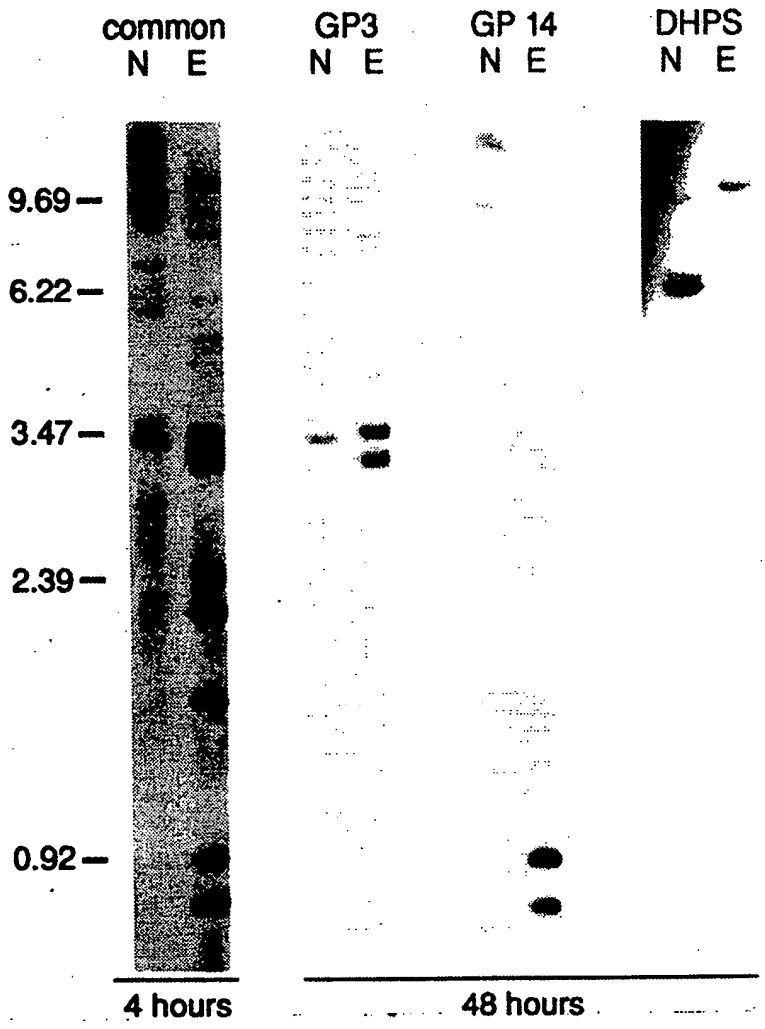
common
N E
GP3
N E
GP 14
N E
DHPS
N E
9.69
6.22
3.47
2.39
0.92
4 hours
48 hours

FIG. 3A

Southern Blot of gp116 (PC5) Hybridized to *P. carinii* Chromosomes Separated by PFGE

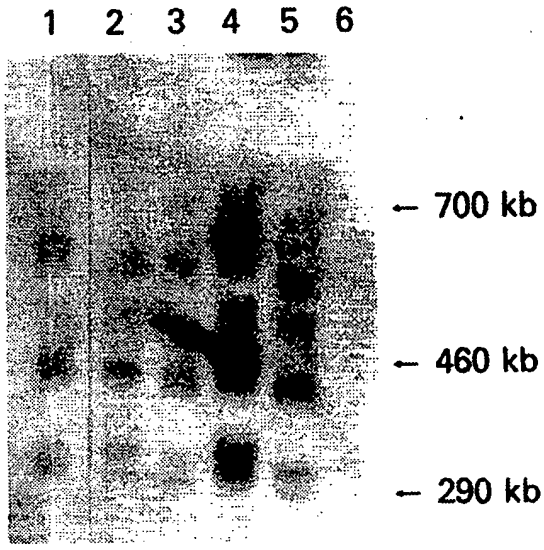

FIG. 3B

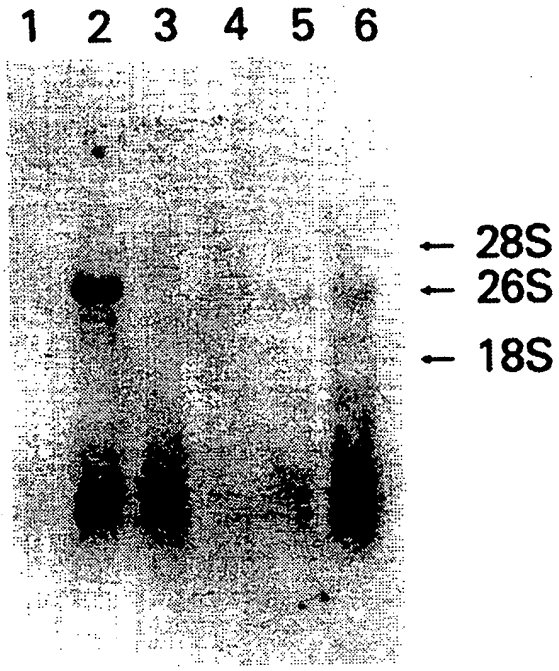

FIG. 3C

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 93/09635

A. CLASSIFICATION OF SUBJECT MATTER
IPC 5 C12N15/31 C07K15/00 G01N33/569 A61K39/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 5 C07K C12N A61K G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	J. CLIN. INVEST. vol. 87, no. 1 , 1991 pages 163 - 170 B. LUNDGREN ET AL. 'Purification and characterisation of a major human P. carinii antigen' see the whole document ---	8-12
X,O	31ST INTERSCIENCE CONFERENCE ON ANTIMICROBIAL AGENTS AND CHEMOTHERAPY PROGRAM ABSTR. vol. 31, no. 0, 1991, page 136 J. FISHMAN ET AL. "Molecular cloning and analysis of genes ..." see the abstract --- -/--	1,2,5-7, 9-12

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

1

Date of the actual completion of the international search	Date of mailing of the international search report
28 February 1994	21 -03- 1994

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	Skelly, J

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	INF. IMMUN. vol. 57, no. 9 , 1989 pages 2149 - 2157 J. RADDING ET AL. 'Identification and isolation of a major cell surface glycoprotein of P. carinii' see the whole document	8,9,11, 12
X	J. PROTOZOOL. vol. 38, no. 6 , 1991 pages 8S - 10S A. SMULIAN ET AL. 'Expression cloning of P. carinii antigens' see the whole document	1,2
X,P	J. BIOL. CHEM. vol. 268, no. 8 , 1993 pages 6034 - 6040 J. KOVACS ET AL. 'Multiple genes encode the major surface glycoprotein of P. carinii' see the whole document	1-12
X,P	WO,A,93 07274 (THE GENERAL HOSPITAL CORPORATION) 15 April 1993 see the whole document	1,2,5-7, 9-12

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US 93/09635

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO-A-9307274	15-04-93	AU-A- 2869192	03-05-93

RECEIVED
JUN 30 1998
KLARQUIST, SPARKMAN, CAMPBELL,
LEIGH & WHINSTON